Fragile Self-Esteem[*]

Botond Kőszegi
Central European University

George Loewenstein
Carnegie Mellon University

Takeshi Murooka
Osaka University

September 7, 2021

Abstract

We develop a model of fragile self-esteem — self-esteem that is vulnerable to objectively unjustified swings — and study its implications for choices that depend on, or are aimed at enhancing or protecting, one's self-view. In our framework, a person's self-esteem is determined by sampling his memories of ego-relevant outcomes in a fashion that in turn depends on how he feels about himself, potentially creating multiple fragile "self-esteem personal equilibria." Self-esteem is especially likely to be fragile, as well as unrealistic in either the positive or the negative direction, if being successful is important to the agent. A person with a low self-view might exert less effort when success is more important. An individual with a high self-view, in contrast, might distort his choices to prevent a collapse in self-esteem, with the distortion being greater if his true ability is lower. We discuss the implications of our results for mental well-being, education, job search, workaholism, and aggression.

[*] We thank Chishio Furukawa, Nicola Gennaioli, Johannes Maier, Tímea Molnár, three anonymous referees, and seminar and conference audiences for insightful comments. Haoxing Ma, Jacob Rothschild, and Greg Volynsky provided excellent research assistance. Kőszegi thanks the European Research Council for financial support under Grant #788918. Murooka thanks financial support from JSPS KAKENHI (JP16K21740, JP18H03640, JP19K01568, JP20K13451).

1 Introduction

The idea that "ego" or self-image is an important determinant of well-being and driver of behavior goes back to Adam Smith,[1] and is increasingly acknowledged in the profession (e.g., Bénabou and Tirole, 2016). In particular, a substantial theoretical literature recognizes that to increase ego utility, anticipatory utility, or internal motivation, or due to information-processing biases, individuals may have unrealistically positive views of their traits or prospects (Carrillo and Mariotti, 2000, Gervais and Odean, 2001, Bénabou and Tirole, 2002, 2011, Prelec and Bodner, 2003, Zábojnik, 2004, Santos-Pinto and Sobel, 2005, Kőszegi, 2006, Grossman and van der Weele, 2017). A corresponding empirical literature has documented exaggerated self-views in many important domains, including IQ (Burks et al., 2013, Charness et al., 2018, Benoît et al., 2015), performance on the job (Malmendier and Tate, 2005, Huffman et al., 2019, Hoffman and Burks, 2020), self-control (e.g., Shui and Ausubel, 2005), health (Oster et al., 2013), and finding a job (Spinnewijn, 2015).

When modeling behavior with unrealistic self-views, economists have uniformly assumed that at any given moment, a person bases his decisions on a single (probabilistic) assessment, taking the same actions as someone with an identical, but realistic assessment. Drawing on an extensive psychology literature documenting that people often hold multiple conflicting views about themselves that can simultaneously influence behavior, this paper formalizes and applies to economics a novel consideration: the *fragility* of beliefs about the self.

The key assumption of our model, consistent with a variety of psychological mechanisms, is that a person's belief about his skill or ability — his self-esteem — depends on his recall and interpretation of memories; but what memories come to his mind and how he interprets them depend, in turn, on his self-esteem. A person experiencing high self-esteem feels good about himself, and is more likely to recall successes and accomplishments, or interpret what he recalls positively; and a person experiencing low self-esteem feels bad about himself, and is more likely to recall failures and blunders, or interpret what he recalls negatively. This reciprocal interdependence

[1] Smith devotes Part III of the Theory of Moral Sentiments (1759) to the "Foundation of Our Judgments Concerning Our Own Sentiments and Conduct, and of the Sense of Duty." He emphasizes the pleasure of "self-approbation," and is careful to distinguish it from social image or approval. For instance, Chapter II is titled "Of the Love of Praise, and of that of Praise-Worthiness; and of the Dread of Blame, and of that of Blame-Worthiness."

creates the potential for multiple "self-esteem personal equilibria" or "SEPs" — levels of self-esteem that generate mental streams of evidence supporting them — with dramatically different, fragile self-assessments. Furthermore, when a person is at one SEP, he consciously or subconsciously understands that ego-shocks might shift him to another SEP. The most common and important situation — one that has been discussed extensively by psychologists and can certainly be recognized by most academics in their colleagues or themselves — is the case of high but fragile self-esteem, a situation we refer to as "insecurity."

We show that the contradictory self-views associated with fragile self-esteem, and especially insecurity, can help to make sense of a range of otherwise perplexing patterns of behavior. A general feature of our results distinct from any other theory is that a person's actions depend on both his self-esteem and — by affecting his level of insecurity — an assessment of his capabilities that would be objectively correct given his experiences. Hence, among individuals with the same high self-esteem, those with lower true ability are especially prone to exhibiting ego-defensive behaviors. 'Workaholics', who often view themselves positively but harbor self-doubts, put extremes of time and effort into work that undermine the quality of their non-work life and even their productivity; people with insecure self-views self-handicap, choosing tasks that are trivially easy or impossibly difficult; and many people with high but fragile self-esteem display physical and emotional forms of aggression in response to the smallest insult; all attempting to maintain a high view of themselves that they deep down appear to recognize is not fully realistic and stable.

We discuss evidence for our central assumption, and present the formal framework outlined above, in Section 2. In Section 3, we establish some basic properties of fragile self-esteem. In many results, a key parameter is the importance of success to the agent, which may depend on both his psychology (how ego-driven he is) and economic stakes, and determines the effect of self-esteem on feelings and therefore memory. While in most existing theories stakes have a positive or no effect on the accuracy of beliefs, in ours they tend to have a negative effect. If success is unimportant, then the agent's self-esteem is close to realistic. If success is sufficiently important, then his self-esteem is fragile and either unrealistically high or unrealistically low. And if success is extremely important, then the possible levels of his self-esteem are essentially independent of true ability.

This multiplicity of SEPs helps to connect the large literature in economics and psychology on unrealistically positive views to a hitherto separate literature in psychology on unrealistically negative views — both observed, typically, in domains of life that are important to the individual. Beyond overconfidence, researchers have documented the prevalence of an "impostor syndrome," whereby a person views himself as *less* capable than objectively justified by his success, and fears that others will soon come to share his perspective (Ferrari and Thompson, 2006). Exactly as in a low fragile SEP, a person suffering from the impostor syndrome may know that his low self-view appears to conflict with available evidence, yet still maintain his self-view.

Our model predicts that if there are multiple SEPs, then the agent re-equilibrates to the same SEP after a small shock to his self-views or feelings about himself, but not after a large shock. This property distinguishes our model from work on biased recall of memories in which cues are exogenous (e.g., Mullainathan, 2002, Bordalo et al., 2020). Indeed, because it will be crucial for behavior, we define the stability — the converse of fragility — of the agent's self-esteem as the smallest shock that can induce a shift to another SEP. We establish connections between the realism of a SEP belief, which is potentially measurable by an observer, and its fragility: the more negative facts there are in the evidentiary base, the more fragile a high SEP and the less fragile a low SEP will be; and if the agent's self-assessment is sufficiently unrealistically positive (respectively negative) relative to his experiences, then he must be in a high (respectively low) fragile SEP. These connections open the possibility for testing our predictions regarding the agent's behavior in a fragile SEP.

In Section 4, we consider the implications of our model for ability-dependent actions. We assume that the effort the agent puts into an activity (e.g., studying or looking for a job) is increasing in his perceived ability, due either to the classical reason that effort and ability are complements, or to the more psychological reason that the cost of effort is decreasing in self-esteem. We show that in such a situation, a person in a low SEP might exert lower effort if success is more important. This can be thought of as a kind of "choking," although a different kind than that studied in the psychology literature: In a person for whom success is especially important, the low-SEP trap is especially severe and therefore demoralizing and discouraging of effort, outweighing the greater importance of success. In contrast, in a high SEP the agent always exerts greater effort if success is more important. While some other models have the implication that incentives can lower effort, they do not predict that this is especially likely to be the case for low-self-esteem individuals.

In Section 5, we consider how the agent might adjust his behavior to influence his self-esteem, focusing on choices that he makes to protect a high SEP by affecting how he feels about himself at the moment. Most straightforwardly, he may exert effort — e.g., preparing sharp rebuttals to anticipated criticism — that directly improves his feelings. Our theory predicts that such efforts increase as ability declines, so that among individuals with the same high self-esteem, it is especially lower-ability individuals who are desperate to maintain a high ego. This prediction contrasts with that of signaling models of self-assessment, in which a higher type always chooses a higher level of the costly signal to increase his self-view.

Another way for the agent to insulate himself from a drop in self-esteem is to avoid shocks to his feelings about himself. Because only a sufficiently large negative shock leads to a collapse in self-esteem, the agent is especially averse to left-skewed shocks, such as that from failing at an easy task. This provides an interpretation of self-handicapping, a widely observed pattern of behavior whereby people introduce obstacles for themselves that they can later blame for their failures (Higgins et al., 2013). Existing work interprets self-handicapping in terms of an information-avoidance motive, but this motive does not robustly predict a disproportionate aversion to left-skewed information, and it should be small for a person who already has plenty of information about himself. Furthermore, the information-avoidance account completely fails to explain the specific circumstances under which self-handicapping has been observed. In typical experiments, subjects are first given surprise positive feedback and are in a position of doubt as to whether it was a fluke, seemingly setting up a high but fragile self-esteem.

In Section 6, we argue that our theoretical framework can help to make sense of a range of phenomena that are or should be of interest to economists. We begin with some insights that relate to findings from the large psychology literature dealing with happiness and well-being. Our model says that memory-based cognitive strategies to improve affect, such as recalling memories of positive experiences, are likely to have only a temporary effect; that very short-lived traumatic experiences can, in contrast, have long-lasting effects; and that the effect of a sequence of shocks can depend on how quickly in succession they occur.

In the rest of our applications, we focus on the behavioral implications of high but fragile self-esteem. A common theme is that, despite their positive immediate self-views, individuals with high self-esteem recognize their vulnerability and distort their choices to avoid "rocking the boat." A 'workaholic' employee may put in unreasonably large amounts of effort, even at a severe cost to other, possibly less ego-related, aspects of life, to guard against the possibility of a tail-spin-inducing setback at work. An unemployed individual may try to avoid being rejected by avoiding job search altogether, or by endlessly searching ads for the appropriate job instead of risking rejections by sending applications. A high-self-esteem student may avoid studying if the occasional failure associated with trying would destroy his high SEP. But, consistent with research documenting the consequences of a 'growth' as compared with a 'fixed' mindset, a student who believes that ability can be developed is at lower risk of such an emotional collapse after a setback. Finally, in accordance with a large psychology literature on aggression, some people with high but fragile self-esteem will lash out when doing so can protect them from a precipitous drop in self-esteem.

As we have emphasized, our paper builds on recent work by economists that has recognized ego or self-worth, as well as people's uncertainty about their own self-worth, as key motives driving human behavior. More generally, our paper belongs to the growing literature that explores the economic consequences of incorrect self-views.[2] Nevertheless, our model's central mechanism based on the two-way interaction between memory and self-esteem is different from others in the literature; and to our knowledge, ours is the first theory in which both objective and subjective assessments influence behavior *at the same time*.[3]

[2] See, e.g., O'Donoghue and Rabin (1999) and Eliaz and Spiegler (2006) regarding naivete about future preferences, and de la Rosa (2011) and Heidhues et al. (2018) regarding overconfidence.

[3] The central result in Oster et al. (2013) that people forego testing for Huntington's disease lest they burst the bubble of overoptimism has a similar flavor, but the formalization is different and specific to medical testing. In the sense of psychologically modeling how self-views come about, our framework is most related to Benabou and Tirole's (2011) self-signaling model of investments in moral identity, in which the agent can perform prosocial actions to influence his future self's beliefs about how altruistic he is. Their model generates a rich set of predictions for behavior depending on the salience of identity, the ambiguity of the situation, and other considerations, that are largely different from ours. Even when Bénabou and Tirole (2011) discuss some of the same phenomena, the mechanism in question is different from ours, so their predictions differ in important specifics. For instance, in their setting a person might work exceedingly hard to signal to himself that work is an important area of his life, whereas in our framework workaholics aim to ensure that ego-destroying performance failures do not occur. Their model predicts that workaholism is most likely for an individual with high economic or social assets, whereas ours says that

The model we propose also builds on theoretical work by economists that examines implications of biased recall of memories (Mullainathan, 2002, Bernheim and Thomadsen, 2005, Bodoh-Creed, 2019, Enke et al., 2019, Bordalo et al., 2020, 2021, Wachter and Kahana, 2021). However, in addition to using a different framework for memory, this prior research does not examine implications for self-esteem. Methodologically, our paper applies the notion of personal equilibrium developed for utility from beliefs (Kőszegi, 2010), and follows previous papers that emphasize the possibility of multiple personal equilibria in other domains (Kőszegi and Rabin, 2006, Spiegler, 2016). Finally, we contribute to the psychology literature by providing a formal model of a phenomenon that psychologists have discussed extensively, but in largely qualitative terms (e.g., Maslow, 1942, Kernis et al., 2005).

2 A Model of Self-Esteem and Memory

In this section, we formulate our model of self-esteem determination.

2.1 Evidence on Central Mechanism

We first summarize evidence for the crucial assumption of our model: that the manner in which a person accesses, interprets, or pays attention to self-relevant memories is influenced by how he feels about himself — which we call his "mood." One primary observation is biases in the retrieval of memories termed "mood-congruent memory" (Isen et al., 1978, Bower, 1981). Psychologists have documented differences in recall across individuals who are or are not subject to mood-related clinical conditions such as depression (Watkins et al., 1992, 1996), within the same individual when in naturally occurring happy or sad states (Mayer et al., 1995), and, perhaps most convincingly, in response to experimentally induced moods (Matt et al., 1992). In one study, positive or negative moods induced via exposure to happy or sad music led to substantial differences in recall of positive and negative life events (Miranda and Kihlstrom, 2005). Another study found that happy and sad people remembered more positive and negative details, respectively, about people they had read about in the past (Forgas and Bower, 1987). Especially relevant to our focus on self-esteem, a

it is most likely for an individual whose self-esteem is unrealistically high.

large number of studies have shown that individuals in negative rather than positive moods are more likely to recall episodes of failure or low task performance relative to episodes of success or high task performance (Blaney, 1986). Other research documents preferential memory for aversive experiences and negative information by depressed individuals (e.g., Watkins et al., 1992, Gilboa-Schechtman et al., 2002, Direnfeld and Roberts, 2006).

Mood-congruent memory is closely related to "mood-congruent judgment" (Mayer et al., 1992, Johnson and Tversky, 1983), whereby mood-congruent information is interpreted to be more valid and relevant than mood-incongruent information. From a theoretical perspective, the notion that a person in a bad mood is more likely to recall failures and the notion that he interprets failures to be more relevant are equivalent, as both increase the weight of failure in his beliefs. Similarly, if a person in a bad mood recalls all his experiences but pays more attention to failures when forming his self-view, then again failures will have an excessive effect. These three mechanisms can occur simultaneously; indeed, Everaert et al. (2014) propose that all three effects are operative in depression.

Mood-congruent memory can be viewed as a special case of the context-dependence or associativeness of memory that has been the focus of some theoretical and empirical work in economics (e.g., Mullainathan, 2002, Bodoh-Creed, 2019, Bordalo et al., 2020, 2021, Enke et al., 2019, Wachter and Kahana, 2021). In this interpretation, a good mood is a context that biases recall toward memories associated with that context — i.e., successes. Relative to existing work on context-dependent memory, our model adds two important ingredients: that the context in question is endogenous to recalled memories, creating a two-way interdependence; and that assessments have direct consequences for utility, introducing a motive to manipulate the process.[4]

[4] Yet other mechanisms may also generate a feedback mechanism similar to that in our model. For instance, anxiety about the potential for poor performance could give rise to poor performance (Ariely et al., 2009), creating a direct link between confidence and performance similar to that in Compte and Postlewaite (2004). And an individual who feels confident may project confidence to others, which can cause those people to respond in a trusting or admiring fashion, thus reinforcing the individual's confidence (Johnson and Fowler, 2011, Johnson et al., 2011, Lamba and Nityananda, 2014). These alternative mechanisms may reinforce the memory-based ones.

2.2 Formal Framework

The agent's perception of his ability is derived from sampling his *evidentiary base* of past payoffs $\{s_1, \cdots, s_n\}$. For simplicity in stating our insights and deriving our results, we assume that each fact in the evidentiary base is binary: $s_i \in \{0, k\}$, where $k > 0$. We interpret an outcome $s_i = 0$ as an instance of failure, and an outcome $s_i = k$ as an instance of success, and define the realistic assessment of ability, $a \in [0, 1]$, as the proportion of successes in the evidentiary base.

In the above framework, we normalize the payoff from failure to 0, and k measures the importance of success *relative to* failure.[5] There may be both psychological and economic determinants of k. Individuals may differ in how much they care about success for personality reasons; some people are simply more 'ego-driven' than others.[6] At the same time, economic stakes, i.e., the extrinsic rewards associated with success, surely also affect k.

The agent's actual self-esteem, the analogue of what researchers in the literature call self-image or confidence, is his perceived ability $\tilde{a}$. We define $\tilde{a}$ as the agent's perceived proportion of successes, which is based on a mood-congruent sampling of his evidentiary base. If his mood is m, then he recalls outcome s_i with probability $g(s_i - m)/\left[\sum_{i'} g(s_{i'} - m)\right]$, where $g(\cdot)$ is a positive-valued function that is single-peaked at and symmetric around zero, and is three times differentiable at non-zero values. We let $E[s|m] = [\sum_i g(s_i - m)s_i]/[\sum_i g(s_i - m)]$ be the agent's average memory about outcomes when he is in mood m. Self-esteem $\tilde{a}$ is then determined according to:

Definition 1 (Self-Esteem Determination)**.** A level of self-esteem $\tilde{a}$ is a self-esteem personal equilibrium (SEP) if it is the proportion of successes the agent recalls in a mood m that is a locally stable solution to $m = E[s|m]$.[7]

[5] More generally, we could assume that $s_i \in \{-fk, (1-f)k\}$ for some $f \in [0, 1]$, where $s_i = -fk$ is an instance of failure, and $s_i = (1-f)k$ is an instance of success. The psychologically realistic level of f depends on the situation at hand. When analyzing traumatic experiences, for instance, $f = 1$ may be most realistic: the agent may find "success" in avoiding a trauma to be a normal, neutral outcome, while evaluating a trauma very negatively. On the other hand, in an economic situation in which the agent works for a reward, such as a bonus, $f = 0$ may be closer to realistic: failure is the status quo, and success is a positive outcome. But since f merely induces a constant shift in outcomes, our results are invariant to it. Hence, we use the normalization $f = 0$ throughout the paper.

[6] In addition to k, individuals may also differ in the standards they apply for considering an outcome a success or a failure. The same objective outcome might be coded as a success by the average person but as a failure by Henry Kissinger (one of our examples for workaholism in Section 6.4).

[7] The notion of local stability we require is the conventional one: that there exists $\Delta > 0$ such that $m' < E[s|m']$ for $m - \Delta < m' < m$ and $m' > E[s|m']$ for $m < m' < m + \Delta$.

Definition 1 is broadly motivated by the following idea. How the agent views himself and feels about himself determines the probabilities with which he recalls outcomes he may have forgotten, or affects the way in which he interprets outcomes he has recalled. These new recollections and interpretations may, in turn, change the agent's self-view and mood. Through this stochastic recall process, the agent eventually converges to a stable, steady-state mood m. If he has experienced mood m for an extended period, then the average memory evoked from the evidentiary base must justify mood m; otherwise, his mood would quickly change. Hence, m must satisfy the fixed-point condition $m = E[s|m]$. And for his mood not to drift away from the fixed point in response to random shocks, the fixed point must be locally stable. Although applied in a different context, these motivations are similar to those in Kőszegi and Rabin (2006) and Kőszegi (2010).

Note that $m = k\tilde{a}$: if success is more important to the agent, then his mood is more sensitive to his self-view. Hence, an equivalent way to define a SEP is as a locally stable solution to the equation $\tilde{a} = \text{Prob}[\text{success}|\tilde{a}]$, where $\text{Prob}[\text{success}|\tilde{a}] = E[s|k\tilde{a}]/k$ is the proportion of successes the agent recalls when he has self-esteem $\tilde{a}$ and therefore mood $m = k\tilde{a}$. Similarly to a SEP level of mood in Definition 1, a SEP level of $\tilde{a}$ gives rise to a proportion of memories of successes that justify $\tilde{a}$. It is often convenient to think in terms of this alternative definition.

We assume for the rest of the paper that any solution to $\tilde{a} = \text{Prob}[\text{success}|\tilde{a}]$ is locally unique.[8] It is then easy to show that:

Fact 1 (Existence)**.** *A SEP exists.*

The essential feature of our framework is that there may be multiple SEPs. To illustrate such multiplicity, we introduce a parametric example that we will use repeatedly in the paper:

Example 1. Let $g(x) = 1/(1 + x^2)$.

Figure 1 illustrates SEP determination in this example when actual ability is $a = 1/2$ and $k = 4$. By definition, the agent's SEPs correspond to points where the "mood-memory curve" $\text{Prob}[\text{success}|\tilde{a}]$ intersects the 45-degree line from above. One intersection is at the true ability $\tilde{a} = 1/2$, but it is not from above (i.e., this solution to $\tilde{a} = \text{Prob}[\text{success}|\tilde{a}]$ is not locally stable). The other two intersections, however, are from above and therefore correspond to SEPs.

[8] This rules out knife-edge situations in which the function $\text{Prob}[\text{success}|\tilde{a}] - \tilde{a}$ of $\tilde{a}$ goes along 0 in a neighborhood.

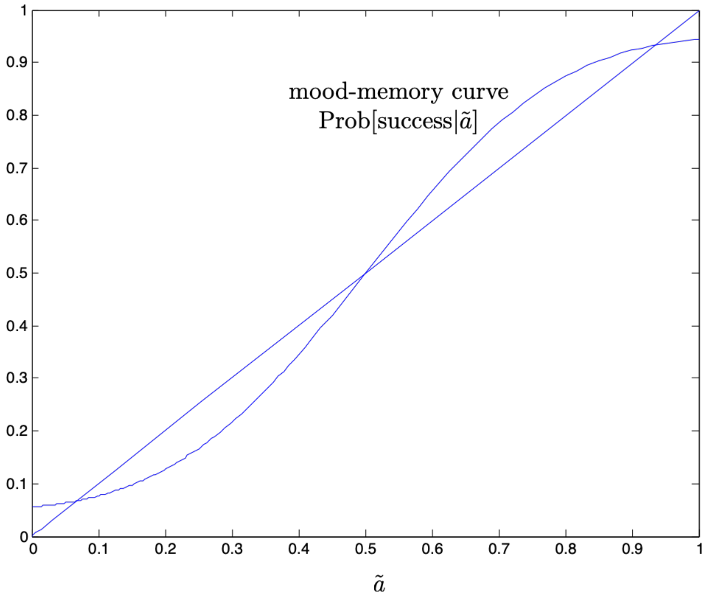

Figure 1: An Example of Multiple SEPs. SEP levels of self-esteem $\tilde{a}$ are stable solutions to Prob[success$|\tilde{a}$] $= \tilde{a}$, which correspond to points where the mood-memory curve intersects the 45-degree line from above.

Based on the recall process that we used to motivate our definition of SEP, we also make a specific assumption about which SEP the agent ends up at. Suppose that the agent starts off at initial self-view $\tilde{a}_0$, or, equivalently, his initial feeling about himself is given by $k\tilde{a}_0$. Then, his self-esteem converges to a SEP according to the following:

Definition 2. Given seed self-view $\tilde{a}_0$, SEP is given by (i) the lowest SEP level $\tilde{a} \geq \tilde{a}_0$ if Prob[success$|\tilde{a}$] $\geq \tilde{a}_0$; and (ii) the highest SEP level $\tilde{a} < \tilde{a}_0$ if Prob[success$|\tilde{a}$] $< \tilde{a}_0$.

Intuitively, if self-esteem $\tilde{a}_0$ induces the agent to recall a higher (respectively lower) proportion of successes than $\tilde{a}_0$, then his self-esteem will drift up (respectively down), so that his self-esteem converges to the lowest SEP above (respectively the highest SEP below) $\tilde{a}_0$.[9] Suppose, for instance, that in Figure 1 the agent starts off with seed self-view $\tilde{a}_0 = 0.4$. He then recalls a lower proportion of successes than 0.4, lowering his self-esteem. This lowers the proportion of successes he recalls, further lowering his self-esteem. Through this process, he converges to the low SEP.

[9] For formal convenience, the definition imposes that if the agent's seed self-view is exactly at an unstable solution to $\tilde{a} =$ Prob[success$|\tilde{a}$], then he ends up at the SEP above it.

This definition allows us to make predictions about how temporary shocks affect the person's self-esteem, and to extend the static model to a dynamic one by assuming that the seed self-view in a period is determined by the previous period's SEP and a shock. The shock in question could be any economic outcome or other new experience that changes how the agent feels — or, equivalently, what he recalls — about himself. Given this perspective, whenever there are multiple SEPs, there is scope for the agent's self-esteem to shift drastically without new information. Suppose, for instance, that in Figure 1 the agent has converged to the high SEP, and now the way he thinks and feels about himself is negatively shocked to determine his new seed self-view. For a small negative shock, the agent returns to the high SEP. For a sufficiently large negative shock, however, his self-esteem collapses to the low SEP. This motivates our final definition:

Definition 3 (Fragility)**.** A person has fragile self-esteem if there are multiple SEPs.

We assume that when the agent is in one SEP, he is (consciously or subconsciously) aware of the other SEPs, and he knows enough about his mood-memory curve to correctly anticipate SEP determination. In light of this sophistication, our previous assumption that the agent's self-esteem is based on a naive average of his memories may appear odd. For instance, if the agent perfectly knew his mood-memory curve and interpreted it the way we as modelers do, then he could infer his true ability a. Nevertheless, we believe that this is not how a person thinks, and more generally our assumptions accurately reflect the underlying psychology. When in a 'down' mood, thinking gloomy thoughts, it is impossible to see oneself in a positive light, even if one recognizes that at other times one has been, and might be again, in a situation characterized by positive thoughts. In such a state, any fact or argument that suggests a higher ability, including the argument that a once-experienced high SEP must reflect a higher proportion of past successes, is also difficult to recall or pay attention to, or is interpreted in a negative light.

For simplicity, we assume in the rest of the paper that there are at most two SEPs. We identify conditions on the primitives for this to be the case:

Proposition 1. *Suppose $g(x) \leq g''(x)$ and $g'(x) \leq g'''(x)$ for all $x \neq 0$. Then, there exist at most two SEPs.*

Roughly speaking, the conditions in Proposition 1 mean that the peak of $g(x)$ is sufficiently sharp: $g(x)$ is sufficiently convex and decreases quickly. For example, $g(x) = e^{-|x|}$ satisfies the conditions. Note also that the conditions are only sufficient, not necessary; Example 1 does not satisfy them, but still has the property that there are at most two SEPs.

As will become clear, a crucial distinction in our model is between the agent's subjective self-view $\tilde{a}$ and his objectively accurate appraisal a. Since a affects the function Prob[success$|\tilde{a}$], it determines the set of SEPs and their fragility. Deviating from previous economic theories of incorrect self-views, therefore, our theory implies that both $\tilde{a}$ and a can influence behavior at any moment in time. Our distinction between $\tilde{a}$ and a parallels the distinction in the psychology literature between explicit self-esteem, which is the level of self-esteem that an individual is consciously aware of, and implicit self-esteem, which is a view of oneself, potentially experienced at an unconscious level, that often encompasses self-doubts (Farnham et al., 1999). In this vocabulary, high fragile self-esteem is characterized by a combination of high explicit and low implicit self-esteem (e.g., Hetts et al., 1999, Koole et al., 2001).

While we focus on self-esteem, virtually the same framework could be applied to other evaluations, such as "How good is my life" (which might correspond closely to "life satisfaction"), "How good are my coworkers," or "How good is the world," that the agent feels strongly about. By the same token, the model could be applied more narrowly to a particular type of self-evaluation — e.g., "How moral am I?" In all of these cases, one can easily imagine that the individual has a bank of memories relevant to the evaluation in question, and that their evaluation affects sampling from the memory bank.

3 Basic Properties of Self-Esteem

In the next three sections, we identify implications of our model. To understand some of the theory's basic logic, and because self-esteem is an important component of utility in itself, we first study self-esteem determination when the agent does not make choices.

Throughout, a central question will be how outcomes depend on the importance of success to the agent (k), and especially what happens when success is very important (k is large). For many of these results, we make the following assumption on $g(\cdot)$:

Assumption 1. $\lim_{x\to+\infty} g(x)x = 0$.

Assumption 1 implies that memories very far from one's current mood have a vanishing effect on self-esteem; that is, the memory process is quite sensitive to mood. Note that Example 1 satisfies Assumption 1.

3.1 Fragility of Self-Esteem

Proposition 2 establishes properties of fragility:

Proposition 2 (Importance and Fragility of Self-Esteem)**.**

I. If $a = 0$ or $a = 1$, then the agent does not have fragile self-esteem.

II. For any $a \in [0, 1]$, if k is sufficiently small, then the agent does not have fragile self-esteem.

III. Suppose Assumption 1 holds. For any $a \in (0, 1)$, if k is sufficiently large, then the agent has fragile self-esteem.

Part I says that if the agent's evidentiary base is completely homogeneous, then he cannot have fragile self-esteem. In this case, he always recalls the same experiences, so there is no scope for mood to influence his memory. Part II says that if success is not that important to the agent, then he again cannot have fragile self-esteem. In this case, he sees little difference between failures and successes, so his mood does not much depend on what he recalls, and the self-reinforcing mechanism of our model is weak. Conversely, Part III says that if the agent considers success to be sufficiently important, then (for interior a) he is guaranteed to have fragile self-esteem. In this case, as the outcomes he recalls change, his mood and therefore his recollections react strongly, making him sensitive to the self-reinforcing mechanism that generates fragility. This means that if a person exhibits the kinds of behaviors below that are associated with fragility more than others, then he reveals that his k is large, i.e., that he is ego-driven.

Figure 2 illustrates Parts II and III by varying k in the example shown in Figure 1. For $k = 0$, Prob[success$|\tilde{a}$] is a horizontal line located at $a = 1/2$, the realistic level of self-esteem. For $k = 1$, the curve is no longer flat, but there is still only one SEP. For higher k, multiple SEPs arise.

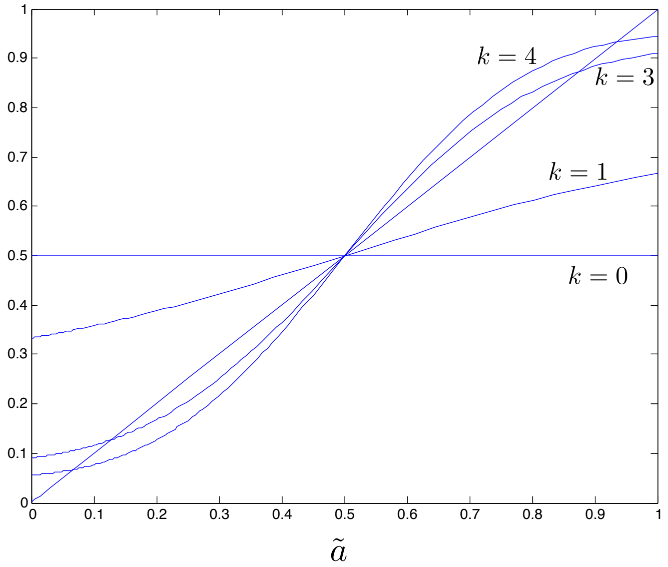

Figure 2: Importance k and Fragility of SEP. A larger k generates a steeper mood-memory curve. For low k, there is a single, realistic SEP. For sufficiently high k, multiple SEPs with unrealistic levels of self-esteem arise.

3.2 Realism of Self-Esteem

We turn to the realism of the agent's self-esteem:

Proposition 3 (Realism of Self-Esteem)**.** *Suppose Assumption 1 holds and $a \in (0, 1)$.*

I. If k is sufficiently large, then there is a SEP with unrealistically high self-assessment ($\tilde{a} > a$) and a SEP with unrealistically low self-assessment ($\tilde{a} < a$), but there is no SEP with realistic self-assessment ($\tilde{a} = a$).

II. The set of SEPs approaches $\{a\}$ as $k \to 0$, and $\{0, 1\}$ as $k \to \infty$.

III. For any abilities a and $a' > a$ and any $g(\cdot)$, if k is sufficiently large, then the corresponding levels of self-esteem $\tilde{a}$ and $\tilde{a}'$ in either the high or low SEP satisfy $\tilde{a} - a > \tilde{a}' - a'$.

Part I says that if ability is sufficiently important, then the agent cannot have a realistic self-assessment, but will have either unrealistically low or unrealistically high self-assessment (see Figure 2 for illustration). When the agent is self-assured, then (for large k) he is much more likely to recall good memories than bad memories, confirming his positive self-view. And when the agent is underconfident, then he is much more likely to recall bad memories than good memories, again confirming his self-view. Part II identifies limit versions of this point: when k is small, a person's mood does not affect the recall process much, so his self-assessment is close to the realistic one; but when k is large, his self-assessment becomes completely dissociated from evidence, becoming almost the highest possible or almost the lowest possible. Finally, Part III implies that when k is sufficiently high and therefore self-esteem is unrealistic, an increase in ability makes high self-esteem more realistic and low self-esteem less realistic.[10]

Proposition 3 has two important implications. First, it establishes that the agent tends to have extreme and unrealistic beliefs when k is high; which, using Proposition 2, also means that he tends to have extreme and unrealistic beliefs when his self-esteem is fragile. In fact, if the agent's self-view is sufficiently unrealistic, he must be in a fragile SEP:

Proposition 4. *If $\tilde{a} - a \leq -1/2$, then the agent is in a low fragile SEP. If $\tilde{a} - a \geq 1/2$, then the agent is in a high fragile SEP.*

Second, our theory ties together hitherto separate bodies of research that have identified systematic biases in individuals' assessments of their own abilities. A large volume of work we have mentioned in the introduction focuses on situations in which people overestimate their own abilities. This overconfidence can be interpreted as a high SEP. While there are previous models of overconfidence, we provide a novel account of it, and identify patterns of beliefs and behavior associated not just with how high beliefs are, but with how *unrealistic* beliefs are. On the other side, a smaller body of research has focused on the "impostor syndrome," whereby an individual interprets his success more negatively than outsiders do and than is realistic, and fears that others will come to share his perspective (Langford and Clance, 1993, Sakulku, 2011, Young, 2011). This underconfidence can be interpreted as a low fragile SEP in which mood-congruent judgment about outcomes the agent readily recalls is operational. We are not aware of a previous model in which the agent both displays underconfidence and understands that his self-view appears to contrast with the available evidence.

[10] At the high SEP, $\tilde{a} - a > \tilde{a}' - a' > 0$, so $\tilde{a}'$ is more realistic, and at the low SEP, $\tilde{a}' - a' < \tilde{a} - a < 0$, so $\tilde{a}$ is more realistic.

Notably, researchers have documented overconfidence and the impostor syndrome for aspects of ability — such as intelligence or work performance — that are important to people. The impostor syndrome, in particular, appears to typically arise in driven, successful individuals with respect to their chosen careers. Our model's prediction that a high k is associated with inaccurate beliefs naturally accounts for this pattern. In contrast, most previous theories do not make the same general prediction. From a classical perspective, one might expect individuals to be best calibrated for the things that are most important to them. Even in existing behavioral-economics models in which an increase in stakes may increase the ego-utility or anticipatory-utility benefit of good beliefs, the agent either responds to the increased incentive to be well-informed or keeps holding the best possible beliefs, so the total effect of stakes on the realism of beliefs is either positive or zero.[11]

3.3 Effect of Shocks

We now explore how vulnerable the agent's self-esteem is to changes. We define the *stability* of a SEP as the infimum of the shock sizes (in absolute value) that can lead the agent to shift to the other SEP. We call a SEP less fragile, or more stable, if its stability is greater. As an illustration, the left panel of Figure 3 varies a in the example in Figure 1. If $a = 1/2$, then the two SEPs are equally fragile; but if $a = 0.4$, then the low SEP is much more stable than the high SEP. Intuitively, if the memories a low-ability but high-SEP agent recalls drop by a bit, the large share of bad experiences he has ignored come flooding back to his memory, setting him off on a tailspin. Generalizing:

Proposition 5. *Suppose there are two SEPs. An increase in a makes the high SEP more stable and the low SEP less stable.*

[11] Some theories of learning with misspecified models do predict a negative effect of stakes on the accuracy of beliefs. Bushong and Gagnon-Bartsch (2021) model an agent who misattributes positive and negative surprises from a experience (e.g., a restaurant meal) to the intrinsic quality of the experience. The greater the stakes, the greater is the surprise on average, so the greater is the misattribution. Relatedly, Heidhues et al. (2018) find that when the loss from choosing the wrong action is greater, the agent's misattribution from her suboptimal choice of action is greater, leading to less accurate beliefs in the long run. These results are driven by completely different mechanisms than ours, and are not applied to beliefs about the self.

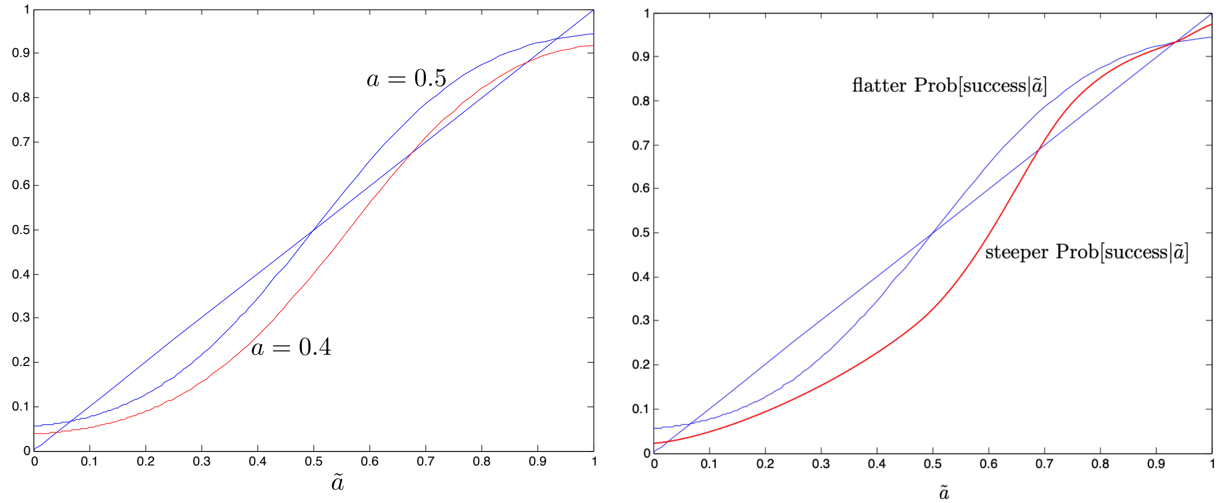

Figure 3: Realism and Fragility of SEPs. In the left figure, a decrease in a makes the high SEP more fragile (Proposition 5). In the right figure, two agents have the same beliefs in a high SEP, but one has a more sensitive memory process (a steeper mood-memory curve Prob[success$|\tilde{a}$]); this agent's belief is less realistic and less stable.

For high k, the combination of Proposition 5 and Proposition 3 Part III imply that among agents who differ only in ability, the fragility of either SEP is increasing in how unrealistic it is. That is, the flimsier is the evidence on which a person's high but fragile self-esteem is based, the more fragile his self-esteem is. A similar comparison can be made between two agents who have different true abilities but identical beliefs, because their memory processes are different. This can happen because the agents have different k or different $g(\cdot)$, or a combination of these factors. Analogously to the single-crossing property in information economics, we can define a memory process as more sensitive if it corresponds to a steeper mood-memory curve, and the two curves cross exactly once (see the right panel in Figure 3). The agent with the more sensitive memory process has lower true ability (at $\tilde{a} = 1/2$, where the weights on failures and successes are the same and hence the mood-memory curve identifies true ability, the red curve is lower) and a more fragile high SEP. This will mean that as distinct from other theories, two agents holding the same belief at the moment of choice may behave differently depending on their true abilities.

The connection between the stability and realism of self-views has an interesting implication:

the fragility of a person's self-view, and its associated behaviors we discuss below, are informative about his ability even if his self-view is less so. As an illustration, consider Example 1, and suppose that k is large. Part II of Proposition 3 implies that for any true ability a, self-esteem $\tilde{a}$ is close to either 0 or 1, rendering the level of self-esteem quite uninformative about ability. But the stabilities of the high SEP and the low SEP are approximately a and $1 - a$, respectively, so the fragility of either SEP is almost perfectly informative about a.[12]

Within the bounds of its stability, a SEP self-assessment is quite impervious to shocks or interventions. Suppose, for instance, that k is large and the agent is in a relatively unrealistic low SEP with $\tilde{a}$ close to zero. This SEP is not very sensitive to a, so providing new information or experiences that change the proportion of successes in the evidentiary base does not by itself change the person's unealistic self-view much. Going further, modest shocks in seed self-esteem do not change the SEP level of self-esteem at all, since the agent simply re-equilibrates to the same SEP. A change in SEP occurs only when the accumulated evidence is overwhelming, and/or the shock to seed self-esteem is large. In this case, the change in self-assessment is drastic. Hence, our model predicts that an individual's self-views may be exceedingly insensitive to information and other manipulations for a while, but excessively sensitive once a critical point is reached. This prediction has a similar flavor to those in models of coarse thinking and inference (Mullainathan, 2000, Mullainathan et al., 2008), in which the agent places objects of interest in rough categories, and therefore underreacts while the category remains the same and overreacts when it changes. But the feature of our model that the shift has a drastic utility impact, and hence the agent is motivated to stay in the high SEP, do not have parallels in models of categorization.

3.4 Measurement

Many of our predictions depend on whether the agent has fragile self-esteem, and, if so, which SEP he is in and how fragile that SEP is. Since the agent's mood-memory curve is not directly

[12] To derive the unstable fixed point of Prob[success$|\tilde{a}$] in Example 1, we solve $\tilde{a} = \text{Prob}[\text{success}|\tilde{a}] = \frac{g(k-k\tilde{a})a}{g(k-k\tilde{a})a+g(k\tilde{a})(1-a)} = \frac{(1+k^2\tilde{a}^2)a}{(1+k^2\tilde{a}^2)a+(1+k^2(1-\tilde{a})^2)(1-a)}$. For a sufficiently large k, this is approximated by $\tilde{a} = \frac{\tilde{a}^2 a}{\tilde{a}^2 a+(1-\tilde{a})^2(1-a)}$, which can be rewritten as $a\tilde{a} + \frac{(1-\tilde{a})^2}{\tilde{a}}(1-a) = a$ or $(1+\frac{1-a}{a})\tilde{a}^2 - (1+2\frac{1-a}{a})\tilde{a} + \frac{1-a}{a} = 0$. By solving this, we obtain that the unstable fixed point is $1 - a$.

observable, the question arises how we can identify these central properties of the agent's self-view.

The results in the previous subsections establish one possible empirical method: by measuring how unrealistic the agent's beliefs are. Section 3.2 shows that an agent with a self-view that is highly unrealistically positive must be in a high fragile SEP, and an agent with a self-view that is highly unrealistically negative must be in a low fragile SEP. Furthermore, Section 3.3 shows that in either SEP, the less realistic the agent's self-view, the less stable it is. So long as an observer has a measure of a and can elicit $\tilde{a}$, therefore, these regularities render our predictions regarding the agent's behavior in a fragile SEP testable.[13]

The psychology literature provides other, completely different approaches to testing our predictions. The most common measure of high fragile self-esteem is a scale measuring narcissism (e.g., Raskin and Terry, 1988, Ames et al., 2006). Narcissists are characterized by a "preoccupation with building, buttressing, and defending... [their] desired self" (Morf and Rhodewalt, 2001), which we can interpret as a high k. Furthermore, according to the Mayo Clinic's website, a narcissist displays "extreme confidence" behind which "lies a fragile self-esteem that is vulnerable to the slightest criticism." In particular, narcissism involves not only a high self-view, but an inflated self-view, and this distinction is crucial for its implications (e.g., Campbell et al., 2004). Given these connections, a reasonable assumption is that people who score higher in narcissism are more likely to be at a high fragile SEP. On the other side, a high level of measured depression, particularly on the part of an accomplished individual, would be indicative of a low fragile SEP. Exactly as in a low fragile SEP, depression is associated with memory, interpretational, and attentional biases that are induced by negative emotions (Gotlib and Joormann, 2010, Everaert et al., 2014).[14]

[13] Of course, in using this methodology some judgment must be made as to how realistic one can expect the agent's self-esteem to be. If the agent is inexperienced, for instance, he may simply need time to arrive at an accurate self-assessment. If he has a lot of experience and is still highly unrealistic, in contrast, he is likely to be misreading the evidence.

[14] Beyond these approaches to testing predictions of the model, it would be useful to develop ways of directly measuring the importance k a person attaches to success. This would make it possible to test our basic prediction that a high k is associated with inaccurate self-views and distorted memories, as well as to test predictions that we develop below concerning the behavior of high-k agents. Crucially, since in our model the agent does not necessarily work hard for something he really values, k cannot be measured straightforwardly using a classical revealed-preference methodology; other approaches, such as survey measures, are necessary.

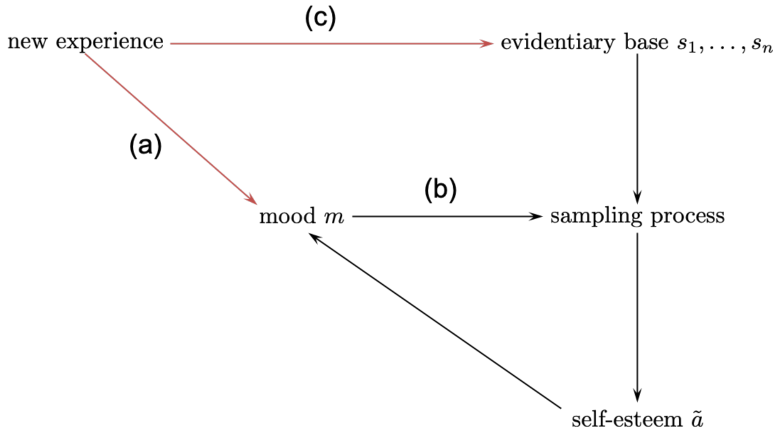

Figure 4: The Effects of a New Experience

3.5 The Utility Impact of a New Experience

Before we turn to how self-esteem affects behavior, we briefly discuss the implications of our model for a person's feelings about himself after experiencing an outcome. The effects can be motivated by a point Thomas Schelling (1987) makes about his self-avowedly irrational self-appraisals after giving a public lecture:[15]

> At my age the statistical record of my performance ought to reflect so many observations of good, poor, and mediocre performance that one more experience at either tail of the distribution could hardly affect a rational self-assessment. I try to remind myself that on those occasions when feedback depresses me; but my welfare function is apparently not constructed that way. It feels to me as if I am taking the audience reaction as evidence, and what makes me feel good or bad is the belief that my average career performance is high or low.

It is plausible that Schelling's thousandth lecture directly influences his mood (path a in Figure 4), but in a classical model this utility impact must be small or short-lived.[16] Our model, in contrast,

[15] A similar observation was made much earlier by William James, the path-breaking American psychologist, who actually coined the term "self-esteem." He wrote that "we ourselves know how the barometer of our self-esteem and confidence rises and falls from one day to another through causes that seem to be visceral and organic rather than rational, and which certainly answer to no corresponding variations in the esteem in which we are held by our friends" (James, 1890, page 307).

[16] If instantaneous utility is determined by a combination of many experiences, then the effect of a single talk is likely to be diluted by other recent experiences and should be small; and if instantaneous utility is dominated by only one or two of the most recent experiences, then the effect of a talk is quickly overtaken by other, newer experiences and is therefore short-lived.

is consistent with Schelling's outsized reaction. The lecture has an indirect effect that acts through the influence of mood on the sampling process from the evidentiary base (path b). Schelling's immediate emotional reaction can constitute a new seed self-view which, if it leads to a shift past the unstable fixed point, can have a dramatic impact on self-esteem. A new experience can also affect utility by becoming an element in the evidentiary base (path c). Even when a new experience does not immediately dislodge the individual from an existing SEP, such a new experience can still change the fragility of that SEP. As Schelling suggests, this last channel is more important for people who have a smaller "statistical record" of experiences.

4 Self-Esteem-Dependent Choices

For the rest of the paper, we analyze the implications of fragile self-esteem for behavior. A person's self-view can affect his actions, and his actions may also affect his self-esteem. While in general both considerations can operate simultaneously, for simplicity of exposition we separate them, considering the former in the current section and the latter in the next section.

We assume that the agent's probability of success depends on his ability a and his effort $e \in [0, 1]$, and that an observation in his evidentiary base as well as his imminent payoff are determined in the following way. Two observable random variables $s_i, s'_i \in \{0, k\}$ are drawn, which equal $k > 0$ with probabilities a and e, respectively. If $s_i = s'_i = k$, then the agent receives a payoff of k; otherwise, he receives a payoff of 0. To analyze choices that are influenced by but do not influence self-esteem, we assume that the agent's evidentiary base codes only the realizations s_i; this amounts to assuming that he can separate the contribution of his ability to performance. His cost of exerting effort level e is $c(e)$, where $c(\cdot)$ is a strictly convex cost of effort with $c(0) = c'(0) = 0$ and $\lim_{e\to 1} c'(e) = \infty$.

Given the above considerations, the agent aims to maximize his perceived intrinsic utility $k\tilde{a}e - c(e)$, where $\tilde{a}$ is determined in a SEP as defined in Section 2. Hence, the agent's effort is strictly increasing in his perceived ability, and it is optimal if and only if his perceived ability equals his true ability.

Three remarks regarding our model are in order. First, a key feature, which derives from the complementarity between ability and effort, is that effort increases in self-esteem. An alternative microfoundation for the same feature is that the agent's energy level and motivation are directly positively affected by self-esteem or mood, such as when a depressed person feels no inclination to do anything productive.[17] Second, we have not included utility from self-esteem in the agent's utility function. In the current model, this would make no difference, as the agent cannot affect his self-esteem. Third, the assumption that k determines both current incentives and the payoff from successes in the evidentiary base amounts to imposing that the strength of incentives is permanent. A temporary increase in incentives does not affect the evidentiary base and therefore should not in itself change the agent's SEP, so our model says that it will increase effort.

We analyze the effect of the importance of success on effort choices:

Proposition 6. *Suppose $a \in (0, 1)$.*

I. If k is sufficiently small, then the agent's effort is strictly increasing in k.

II. Suppose Assumption 1 holds. If k is sufficiently large and the agent is in the low SEP, then his effort can be strictly decreasing in k. In particular, his effort approaches 0 as $k \to \infty$.

III. Suppose Assumption 1 holds and $g(x)x$ is decreasing for sufficiently large x. If k is sufficiently large and the agent is in the high SEP, then his effort is strictly increasing in k and approaches 1 as $k \to \infty$.

As we have seen in Proposition 2, when k is low, the agent's SEP is unique. Though his SEP can be either increasing or decreasing in k in this case, $k\tilde{a}$ is strictly increasing in k. Intuitively, if k is sufficiently small, then the agent's recall process is not much affected by k, so an increase in k does not change his self-esteem much. Then, the direct effect of increasing k dominates, and the agent reacts in a classical way to incentives.

Proposition 2 also implies, however, that if k is sufficiently large, then the agent's self-esteem is fragile. In this range, an increase in k can *lower* the effort of an agent in the low SEP. To see this, recall our Example 1 with $a = 1/2$. If $k > 2$, there are two SEPs, $\tilde{a}^{low} = \frac{1}{2}\left(1 - \sqrt{1 - \frac{4}{k^2}}\ \right)$

[17] Formally, suppose that (i) the agent gets a payoff of k if *either* s_i or s'_i equals k, so that there is no complementarity between ability and effort; and (ii) the cost of effort is $c(\tilde{a}, e)$, with c and c_e both decreasing in $\tilde{a}$. Hence, perceived intrinsic utility is $k(\tilde{a} + e) - c(\tilde{a}, e)$. The predictions of this formulation differ in only two ways from those of the formulation above. First, in this alternative version effort is always optimal, whereas in the above version it is often too high or too low. Second, in this version the agent's welfare is always higher in the high than in the low SEP, whereas in the above version it might not if his action is more distorted in the high than in the low SEP.

and $\tilde{a}^{high} = \frac{1}{2}\left(1 + \sqrt{1 - \frac{4}{k^2}}\,\right)$, and it is easy to verify that in the low SEP, the agent's expected achievement $k\tilde{a}^{low}e$ is strictly decreasing in k. Intuitively, if success is really important, then being pessimistic about one's ability feels far worse than being optimistic, and successes are very happy memories relative to failures. As a result, successes are especially difficult to recall when down, exacerbating the agent's negative self-assessment. This low-SEP trap can be so severe that it outweighs the classical effect of incentives, and means that the agent does not work hard for the success he highly values. While the agent's effort may not be monotonically decreasing in k in general, under Assumption 1 his effort approaches 0 as $k \to \infty$.

When applied to a single individual across domains, the above logic says that low-self-esteem individuals may perform the worst for tasks on which success is most important. In this sense, our model predicts a novel mechanism for choking that is completely different from reasons discussed in the psychology literature, such as supra-optimal levels of arousal, the diversion of attention from the task to the consequences of performance, and the involuntary invocation of conscious processing in place of superior automatic processes (Baumeister 1984, Baumeister and Showers 1986; and in economics, Ariely et al. 2009).

In contrast, Part III says that, with the additional condition on $g(\cdot)$, the possibility of choking applies only to agents in a low SEP, so that our framework predicts that only low-self-esteem individuals can react in a perverse way to incentives. When k is higher, an agent in a high SEP is less likely to recall bad facts about himself, increasing his self-confidence. This reinforces the classical effect of incentives.

Existing models of self-image or social image are also consistent with a negative effect of incentives. In Bénabou and Tirole (2006) and Kőszegi and Li (2008), for instance, an increase in incentives raises doubts as to why a person might perform well, lowering the image benefit of effort. A distinctive prediction of our model, however, is that the negative effect of incentives is limited to individuals with an unrealistically low SEP.

5 Self-Esteem-Influencing Choices

In this section, we analyze actions that are motivated by a desire to either maintain or increase self-esteem, focusing on efforts to remain at a high SEP. Our model says that the agent may improve his self-esteem by taking actions to affect how he feels about himself, thereby manipulating his seed self-view, and by trying to add favorable new observations to his evidentiary base. For many or most of the important aspects of self-esteem, however, most people have an extensive evidentiary base, so adding further observations is likely to have a small effect in itself. Hence, we restrict attention to the first channel, manipulating momentary feelings about oneself. This is also useful from the methodological vantage point of distinguishing our model from information-acquisition-based theories.

In general, we think of the agent as starting off with self-esteem $\tilde{a}_{-1}$, which we will typically take to be a SEP. The agent then chooses an action e, which produces some material costs and benefits. Each action is also associated with a distribution of shocks to self-esteem, after which the agent converges to a SEP according to our model. The agent's goal is to maximize the sum of his expected self-esteem and material utility.[18] Within this general framework, we analyze two specific ways of managing self-esteem shocks: increasing his feelings about himself directly, and avoiding random shocks to his self-views.

5.1 Improving Feelings about Oneself

Consider a situation in which an individual faces the prospect of a shock to his self-esteem, which, he recognizes, could establish a new seed self-view. The individual can, however, take actions that mitigate the effect of the shock. An employee who proposes a new idea for consideration by his team, for example, might fear that others will identify shortcomings in, and criticize, his idea, which could plunge him into a low SEP. To prevent this from occurring, the employee can take extra effort to reduce the severity of his emotional shock. Some of these steps could be directed at colleagues' possible reactions — e.g., preparing irrefutable rebuttals to possible objections, or

[18] Formally, we assume that the agent prefers high self-esteem because it directly increases his utility. He may also prefer high self-esteem because it increases motivation (as in Carrillo and Mariotti, 2000, Bénabou and Tirole, 2002) or has a strategic advantage. In a reduced form, such considerations can be included in the utility from self-esteem.

intimidating some of his co-workers. Other steps might be directed at his emotions — e.g., taking mental steps to dampen emotional swings.

Formally, the agent starts off with self-view $\tilde{a}_{-1}$, and receives a shock $\epsilon \in \mathbb{R}$ to his mood. After observing the shock, he can take an action $e \geq 0$ at cost $c(e)$, where $c(\cdot)$ is strictly convex with $c(0) = c'(0) = 0$. The shock and effort determine his seed self-view according to $\tilde{a}_0 = \tilde{a}_{-1} + \epsilon + e$. Returning to the example of the employee, ϵ is his perception of the reactions he will face, and e is the effort he makes to refute any objections.[19] We could assume that the effect of effort is probabilistic; this would not change the qualitative conclusions regarding the agent's willingness to exert effort, but it would mean (realistically) that the agent cannot guarantee himself a high SEP. Finally, the agent's SEP is determined as in our basic model with seed self-view $\tilde{a}_0$. Since our primary goal is to analyze actions motivated by the desire to increase self-esteem, we assume that the agent's goal is to maximize his resulting level of self-esteem minus the cost of effort, $\tilde{a} - c(e)$.

Proposition 7.

I. If the agent does not have fragile self-esteem, then $e = 0$.

II. If the agent has fragile self-esteem, then there exist $\underline{\epsilon}$ and $\overline{\epsilon}$ such that $e = \overline{\epsilon} - \epsilon$ for $\underline{\epsilon} < \epsilon < \overline{\epsilon}$, and $e = 0$ otherwise. For a fixed $\tilde{a}_{-1}$, or if $\tilde{a}_{-1}$ equals either the high or the low SEP, $\overline{\epsilon}$ is strictly decreasing in a.

Part I says that if the SEP is unique, then the agent does not exert mood-increasing effort. In this case, he cannot influence his SEP level of self-esteem. More interestingly, Part II says that if the agent's self-esteem is fragile, then for a range of shocks he exerts positive effort that is (for fixed or SEP starting self-esteem $\tilde{a}_{-1}$) strictly decreasing in his true ability. Conversely, the more we see an agent with a given self-view engaging in ego-defensive behavior, the lower, we can conclude, is his ability. Intuitively, whenever the agent faces a seed self-view that would result in the low SEP, he can increase his self-esteem by improving his seed self-view to the unstable fixed point, ending up at the high SEP. If the shock would in itself put him sufficiently close to the unstable fixed

[19] In the most straightforward interpretation, the shock ϵ is a current challenge to the agent's self-esteem, which he can mitigate through his effort. Alternatively, as in the above example, the shock could be a future challenge, with him taking a preemptive action to change the effect of the challenge.

point, the discrete improvement in self-esteem is worth the effort cost. Furthermore, lower ability makes the high SEP less stable, making it harder to manipulate oneself to the high SEP.

The result that effort could be decreasing in ability runs counter to the predictions of signaling models of self-assessment, including Bodner and Prelec (1996), Prelec and Bodner (2003), Bénabou and Tirole (2011), and Grossman and van der Weele (2017). A basic property of signaling models is that a higher type chooses a higher level of the costly signal — here, effort — than a lower type. More fundamentally, no other model makes the prediction that behavior depends not only on how high, but also on how *unrealistically* high a person's self-view is.

From the perspective of the current model, the agent has an incentive to increase his seed self-view in both SEPs: if in the high SEP, he may exert effort to maintain it, and if in the low SEP, he may exert effort to shift to the high SEP. The former type of behavior is more likely to occur than the latter, however, in natural variants in which the agent's effort and self-view are complements either for classical technological reasons or because the cost of effort under low self-esteem is high (see Section 4 for a discussion). Furthermore, in the conclusion we provide another conjecture as to why it might be difficult to get out of a low SEP.

5.2 Avoiding Shocks

The agent may also protect a high self-esteem by selecting the types of activities to engage in. To analyze this motive, we endogenize mood shocks based — as in models of disappointment aversion (Bell, 1985, Loomes and Sugden, 1986, Kőszegi and Rabin, 2007) — on how an outcome differs from the agent's expectations; other considerations of what affects the agent's feelings about himself can readily be incorporated into our model as well, but we do not do so here.[20] Suppose that the agent engages in a task that yields success with probability p and failure with probability $1-p$. He then experiences a mood shock of $(1-p)r$ if he succeeds and a mood shock of $-pr$ if he fails, where $r \geq 0$ measures his emotional sensitivity to the outcome. If he does not engage in the task, then he is not subject to shocks. The agent starts off in a high fragile SEP, and, after his new seed self-view

[20] Among many other sources of ego-relevant shocks, one important category is probably social comparisons. It may be the case, for instance, that observing the successes of close acquaintances generates negative shocks. If so, the agent can partly manage such shocks by choosing the types of people he associates with.

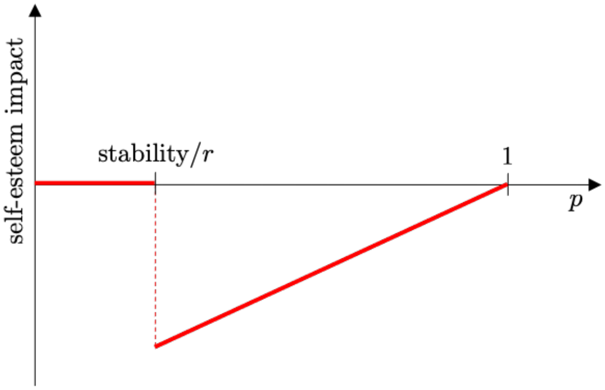

Figure 5: The Expected Self-Esteem Impact of Engaging in a Task with Success Probability p, when starting from the high SEP. If p is low, failure is a small disappointment, so self-esteem returns to the high SEP. If p is sufficiently high yet failure occurs, self-esteem drops.

is determined, he converges to a SEP according to our model.

If r is lower than the stability of the high SEP, then failure cannot put the agent's self-esteem below the unstable fixed point of his mood-memory curve Prob[success$|\tilde{a}$], so he always returns to the high SEP. Suppose, therefore, that r is strictly greater than the stability of the high SEP. Figure 5 shows how engaging in this task affects the agent's expected self-esteem. If success is very unlikely, then failure is not much of a disappointment, so the agent's mood does not drop much, and he re-equilibrates to the high SEP. If success is sufficiently likely, however, a failure is so disappointing that self-esteem collapses, creating a discontinuous negative impact on self-esteem. For still higher probabilities of success, failure still leads to a collapse in self-esteem, but this is now less likely, so the expected impact is less severe. Finally, if success is guaranteed, then again the task has no impact on self-esteem.

These results imply distortions in the agent's behavior relative to what is materially optimal. In particular, consider different tasks with the same r and the same extrinsic net-of-cost material payoffs. Among these tasks, the agent prefers tasks with a high or low probability of success over tasks with an intermediate probability of success.[21] Furthermore, the agent has a particular preference for right-skewed risk over left-skewed risk: unless success is guaranteed, he likes low

[21] While this is optimal in our model, in natural extensions that incorporate present bias, it means underinvestment in long-term objectives. See our discussion at the beginning of Section 6.

success probabilities to high success probabilities.

These predictions are consistent with self-handicapping, a widely documented phenomenon in which an individual chooses a task that is inordinately difficult, i.e., has a low p, or a task that is extremely easy, i.e., has a high p. Both psychologists (e.g., Berglas and Jones, 1978, Greenberg, 1985) and economists (Bénabou and Tirole, 2002, Kőszegi, 2006, Kőszegi and Rabin, 2009) can explain self-handicapping in terms of avoidance of the drop in beliefs from negative information: failure is meaningless with an impossible task and unlikely with an easy task. Our explanation based on negative mood shocks is related to these accounts, but moves beyond them in two ways.

First, self-handicapping is not necessarily about information avoidance. Since most people have plenty of information about themselves, a minor piece of news is unlikely to affect their beliefs by more than a trivial amount, so a purely information-based account should not predict a strong motivation to self-handicap. To the extent that a bad talk, a critical observation by a friend, or another unfavorable signal can affect the agent's recall of memories, and threatens to perturb them from a high to a low SEP, however, a seemingly outsized motive to avoid the signal makes perfect sense.

Second, our model gives a fuller explanation than previous accounts for the circumstances under which self-handicapping has been observed. Experiments typically start by giving subjects unexpected favorable feedback, putting them into doubt as to whether this will be repeated. In our interpretation, this prelude induces a high fragile SEP with a large discrepancy between $\tilde{a}$ and a, exactly the situation in which our model predicts self-handicapping is most likely to occur.[22]

6 Applications

In this section, we outline a number of applications of our framework. Note that while our model assumes utility maximization, for interpreting behavior in some of these applications it is important

[22] While our discussion has assumed that the source of uncertainty in the agent's mood is external, it could also be internal. Suppose, for instance, that an individual at a high SEP receives ambiguous feedback from a colleague. As he tries to decipher the feedback, he may vacillate between positive and negative interpretations, generating mood swings. The effect of these swings can be asymmetric. If the agent contemplates a positive interpretation and thus receives a positive mood shock, his self-esteem re-equilibrates to the same level. But if he contemplates a negative interpretation and thus receives a negative mood shock, his self-esteem can drop precipitously. As a result, a person with high fragile self-esteem may be as sensitive to ambiguous feedback as to clearly negative feedback.

to recognize that people are present-biased (e.g., Laibson, 1997, O'Donoghue and Rabin, 1999). Self-esteem is in a significant part a short-term consideration — it has immediate and potentially large effects on mood — while some of the negative consequences of the behaviors that fragile self-esteem produces are more long-term — e.g., personal stagnation due to self-handicapping, family problems resulting from workaholism, and eventual morbidity or mortality resulting from reciprocated aggression. Hence, in attending to self-esteem, present-biased people fail to maximize long-run utility.

6.1 Mental Well-Being

We begin with implications of our model for mental well-being. Our prediction that self-esteem re-equilibrates to the same SEP for sufficiently small shocks but drastically changes for sufficiently large shocks helps to make sense of three disparate but central phenomena that have not been explicitly discussed in existing papers.

First, our model identifies the limits of cognitive strategies, such as consciously trying to recall good memories or being grateful for good things in one's life, that attempt to improve affect. Such strategies are central, for instance, in self- and therapy-induced attempts to escape depression.[23] These strategies presumably evoke some successes from the agent's evidentiary base, or induce him to interpret memories more positively, so they indeed make him feel better. Given, however, that such manipulations are unlikely to have dramatic effects on momentary feelings about oneself, the agent's mood-dependent recall and interpretation of other memories means that his self-esteem returns to the same SEP, limiting the duration and magnitude of the manipulations. Roughly consistent with such a limitation, Josephson (1996) shows that depressed subjects are less successful than non-depressed subjects in shifting toward the retrieval of positive memories (see also Heimpel et al., 2002).

Second, our model helps to understand post-traumatic stress disorder (PTSD), in which a traumatic event — even if it is extremely short-lived and hence should carry little weight under the conventional calculus of discounted utility — has catastrophic long-run impact on a person. In

[23] Forgas (2017, page 96), for example, writes that "sad people eventually may escape the vicious circle of focusing on and remembering negative information by means of deliberately employing mood-incongruent attention and memory."

our interpretation, a trauma plays a dual role. Proposition 2 implies that a lone traumatic event in the evidentiary base can ensure that the agent's evaluation of himself — or, more generally interpreted, his life — is fragile, even if all his other memories are good.[24] Furthermore, the large emotional impact of a trauma can drive the individual into the low SEP, ensuring that he succumbs to fragility. When such collapses occur, memories of the traumatic event dominate his thinking, in effect damning him to relive the experience. Such an account is consistent with a range of research documenting selective attention and memory in victims of PTSD (see, e.g., Buckley et al., 2000).[25]

Third, our framework has implications for the consequences of a sequence of shocks to self-esteem. If adjustments to shocks take time, then the impact of such a sequence can depend on whether the shocks occur one after the other in quick succession or separated by a substantial time delay. An individual who experienced one shock each month, for example, might have sufficient time to return to the high SEP prior to receiving each subsequent shock, leaving him at the high SEP at the end of the sequence. In contrast, an individual at a high SEP who experienced a series of negative shocks in a short period of time would be less likely to have time to re-equilibrate between shocks, and, as a result, would be more likely to end up at the low SEP. This implies that an individual at a high SEP will be reluctant to experience a series of events with unpredictable consequences for mood in close succession, but will be less reluctant if the events are separated by time. Although we are not aware of direct tests of these predictions in the domain of self-esteem, closely related logic is at play in psychology accounts of stress.[26]

[24] Formally, we can capture this situation by assuming that k is large, and there is a single 0 in an evidentiary base otherwise composed entirely of k's. (An equivalent model arises if there is a single terrible outcome $-k$ among 0's. See Footnote 5.) In combination, Parts I and III of Proposition 2 imply that if k is sufficiently large, then the agent's self-esteem is fragile with the trauma, but not without the trauma.

[25] Accordingly, the first 4 items in the "PTSD checklist" (Blanchard et al., 1996) all deal with selective recall of memories: intrusive recollections, flashbacks, being upset by reminders, and experiencing distressing dreams.

[26] The dominant account of stress in the psychology literature, for example, assumes that stress occurs when the challenges facing an individual exceed the protective capabilities of the 'coping' resources that the individual can marshal (Lazarus and Folkman, 1984). There is also some support for the idea that past stressors can cumulate and interfere with coping (e.g., Brewin et al., 2000, Turner et al., 1995). One important study (Bonanno et al., 2007) reported results from a phone study conducted in the New York City area after the September 11, 2001, terrorist attack. People who reported post-traumatic stress disorder symptoms (as compared with those who did not) differed on a number of dimensions, including demographic characteristics, close connection to the traumatic events, and, most important for this prediction, the occurrence of recent life stresses.

6.2 Education

The rate of return on education is high for many who choose not to continue their schooling or put minimal effort into studying. There are two main existing interpretations of this underinvestment, lack of information and underappreciation of future rewards, but neither receives strong support from the empirical literature.[27]

We provide an alternative, or at least additional, interpretation. In our model, both the low and the high SEP has its own unique trap for students. On the one hand, interpreting the effort choice in Section 4 as studying, our analysis implies that students in the low SEP study too little. This is similar to previous accounts in which students base decisions on incorrect information, although it qualifies these accounts in relevant ways. Our arguments in Section 3.3 imply that — consistent with the evidence — investment may not increase much in response to simple informational interventions. Furthermore, Proposition 6's insight that, for someone at the low SEP, effort can be decreasing in stakes can help to explain why students underinvest in something so important.

On the other hand, a low-performing student who manages to nevertheless attain a high SEP may not study because his SEP is quite fragile, and studying would expose him to the risk of a collapse in self-esteem. Even if the student studies, he might choose tasks that are too difficult (i.e., have a low p) or too easy (i.e., have a high p), to the detriment of improvement. While we do not know how important this consideration is in educational underinvestment, one prominent set of recent findings does seem supportive. Researchers distinguish between a 'fixed mindset', which sees intellectual ability as innate and immutable, and a 'growth mindset', which sees intelligence as a capability that can be developed. A large body of research finds that those with growth mindsets fare better academically than those with fixed mindsets (e.g., Uluduz and Gunbayi, 2018).[28]

[27] The first perspective attributes underinvestment to a lack of information about the returns to education. This has led to interventions that provide better information, but such interventions have yielded mixed results at best: some studies, almost exclusively in developing countries, observe beneficial effects on outcomes (Jensen, 2010, Avitabile and De Hoyos, 2018), but others do not (e.g., Fryer Jr., 2013, McGuigan et al., 2016). The second perspective, dominant in the behavioral economics literature, is that, as a result of steep time preferences, students are not sufficiently motivated by long-term rewards such as a successful career in middle age (Lavecchia et al., 2016). This view has led to a variety of interventions that involve providing immediate incentives to compensate for the weak motivational impact of distant benefits. But such interventions have yielded almost uniformly disappointing results (Fryer Jr., 2011, Leuven et al., 2010, Bettinger, 2012, Lavecchia et al., 2016).

[28] In one of the few studies examining these issues conducted by economists, Alan et al. (2019) randomly assigned students in some elementary schools in Turkey to receive an intervention designed to induce "grit." As described by

Conventional logic largely accounts for this phenomenon: students who believe that there are high returns to effort naturally put in greater effort. But there is an aspect of the evidence that existing models do not account for. Studies have found that the benefits of a growth mindset are especially pronounced when students are confronted with setbacks such as a poor grade on a test (Aditomo, 2015, Alan et al., 2019): those with fixed mindsets tend to take a failure to heart and become discouraged, whereas those with growth mindsets tend to persist. In existing models, it is difficult to imagine that one bad grade on a midterm or one snag in understanding the material would drastically change a student's optimal study effort, so significant setback-related discouragement should not occur.

In contrast, a natural version of our model provides a compelling account. In Appendix A, we consider a situation in which a student's self-esteem is based not on his beliefs about his innate ability, but on his beliefs about his future skills, and he can develop his skills by exerting effort. We show that under reasonable conditions, regular studying is easier to sustain for a student with a growth mindset than for a student with a fixed mindset. So long as he studies, a student with a growth mindset believes that he will improve significantly, so his mood-memory curve lies above that of a student with fixed mindset. Analogously to Proposition 5, therefore, he has higher and more stable self-esteem at the high SEP, which is more robust to any emotional down-swings that studying may cause.

6.3 Job Search

Unemployment is one of the most severe and reliable causes of unhappiness, with an impact that is over five times as large as the effect of the income loss itself (Helliwell and Huang 2014; see also Clark and Oswald 1994, Winkelmann and Winkelmann 1998). Yet, the unemployed do little to alleviate this aversive state: they devote surprisingly little time to job search. For example, international comparisons by Krueger and Mueller (2010, 2012) found that unemployed individuals'

the researchers, the intervention "aims to positively influence children's beliefs about the malleability of ability and the productivity of effort in the skill accumulation process, and thereby induce gritty behavior. The program exposes children to a world view in which ability, rather than being innately fixed, can be developed through sustained, goal-oriented effort. The core message is to highlight the role of effort in the skill accumulation process and thereby in achievement, and to discourage students from interpreting early setbacks and failures as evidence for a lack of innate ability." The program yielded a variety of substantial benefits, including raising test scores.

average search time is 3-30 minutes per day, and 80-95% do not search at all. A substantial literature on dropout from the labor market — including from job search — interprets these patterns by suggesting that many unemployed individuals are "discouraged," underestimating their own value as workers and likelihood of finding work if they tried (Goldsmith et al., 1996, Bjørnstad, 2006, Feather, 2012). In our framework, this interpretation corresponds closely to the idea that unemployed individuals are typically at low SEPs.

Another empirical regularity, however, does not sit comfortably with the picture of the pessimistic unemployed. The average unemployed person is in fact *overoptimistic* about finding a job, with the actual average jobless duration being over three times the expected one (Spinnewijn, 2015). One contributing factor to the pattern of overoptimism with little search could be present-biased preferences (e.g., DellaVigna and Paserman, 2005): the unemployed put a large weight on the effort cost of search relative to its benefits, and naively believe that they will be more forward-looking in the future. Yet given how unpleasant unemployment is, and the fact that a little bit of effort is unlikely to be very costly, present bias does not appear to provide an adequate explanation in itself. Another possibility is that the unemployed, while optimistic about their job-finding rates, are pessimistic about the effect of search effort on this rate (what Spinnewijn, 2013, 2015, refers to as "control pessimism"). But it seems unlikely that the unemployed would believe that almost no search optimally balances the effort costs and job-finding benefits of search.

A different possible interpretation of overoptimism about job finding suggested by our model is that a meaningful subsample of the unemployed are at a high SEP. Analogously to education, those who manage to achieve a high SEP despite being unemployed may be afraid to search because their SEP is quite fragile. Searching for a job but not finding one, especially if there is explicit rejection (e.g., being told that one is not qualified for a specific job) can risk a negative emotional shock that not searching for a job, and hence not finding one, does not. In this situation, in a pattern reminiscent of self-handicapping, the individual might not engage in job search — even though he actually overestimates the return to effort.[29] Hence, our model suggests that thinking about the

[29] As with self-handicapping in general, this behavior is arguably not just about information avoidance, but also about avoidance of negative emotional shocks. An unemployed person who, despite not searching, expects a job soon (in 6.8 weeks in Spinnewijn's data) must also expect a lot of information about his qualifications. It is unclear whether searching little lowers the amount of information he will receive.

implications of biased beliefs based on a reduced-form model — such as optimism and pessimism of an unmodeled nature — might sometimes be misleading.

Although we have not found evidence on the finer details of real-life job-search strategies, our model also suggests other possible outcomes. Beyond not searching at all, an unemployed person might distort his choices toward specific types of search activities, such as sifting through want ads or sending out resumes rather than making phone calls or interviewing for jobs, that are less likely to succeed but also less likely to result in painful explicit rejection (i.e., that have a low p). Finally, in line with our discussion in Section 6.1, an individual might be willing to search (and be rejected) a single, or only a limited number of times within a given period, but after reaching some threshold might choose to take a break — providing time to re-equilibrate to the high SEP — before searching further.

6.4 Workaholism

The term "workaholism" was coined by psychologist Wayne Oates in a 1968 essay in which he confessed to being subject to the syndrome. Oates identified workaholism as a disorder akin to substance abuse, and contemporary definitions of workaholism often include components that overlap with definitions of addiction, such as "working beyond what is reasonably expected of the worker ... despite the potential for negative consequences (e.g., marital issues)" (Clark, 2016). Studies of workaholism have identified numerous negative consequences of the syndrome, including job stress, work-life conflict, and burnout.

Consistent with Oates's view, one possible explanation for workaholism might come from economic models of addiction (e.g., Becker and Murphy, 1988, Orphanides and Zervos, 1995).[30] Our framework identifies an additional plausible explanation: workaholism is a self-esteem-influencing choice of the type in Section 5. Individuals who are at high but fragile SEPs may exert effort that guarantees professional success — i.e., it compensates for a potential negative shock ϵ in the notation of Section 5.1 or has a high p in the notation of Section 5.2 — and thereby helps maintain

[30] In such models, the more one engages in the addictive activity, the higher are the relative returns to the activity. This description does, in fact, seem to apply to work. For example, developing friendships takes time, and time spent at work could very plausibly detract from friendships, family life, and other forms of social capital.

their high SEPs. Since fragility in our framework is exacerbated by low ability, our model predicts that among people with similar self-esteem, the more extreme workaholics would be, on average, of lower ability.

Although further empirical research is necessary, a convergent set of suggestive evidence connects our model to workaholism. Studies have found that narcissism, which is an indicator of high fragile self-esteem, is indeed positively related to workaholism (Clark et al., 2010, Andreassen et al., 2012). Furthermore, a recurring finding from research on workaholism is that the long work hours are not associated with high individual productivity (Clark et al., 2016). This is consistent with our prediction that workaholics tend to have lower ability, as well as the prediction that they pursue an inefficient work strategy.

The connection of workaholism to insecurity also accords with the conclusions of qualitative observers. A USAToday article ("Could Insecurity Be the Secret to CEOs' Success?", Jones, 2007) cites CEOs, star athletes, and other major figures who attribute their success to insecurity. In one of many examples, a CEO remarks that "[w]e are all insecure ... and drive very hard to compensate and to prove we are better than the rest." Similarly, a remarkably high fraction of biographers of powerful and influential people make note of the insecurity of their subjects. Robert Caro (2011), for example, repeatedly refers to Lyndon Johnson's insecurity. Ron Chernow (2016, page 145) describes Alexander Hamilton as "a mass of insecurities that he usually kept well hidden." And Walter Isaacson (2005, page 162) describes Kissinger as an "odd mixture of ego and insecurity." The prevalence of insecurity among the rich, powerful and famous might seem surprising given their extreme levels of success by conventional measures, including having biographies written about them, but our analysis suggests that it should not be: insecurity is exactly what drives them to succeed — what gives them such a profound need to prove themselves.[31]

[31] Our model implies that these insecure successful individuals have abilities that are lower than their self-views. But as we have mentioned in Footnote 6, it is likely that for such individuals the standard for success — and therefore the standard for having high ability — is extremely high. Hence, their true ability, as well as the low SEP to which they fear their self-assessment will drop, can still be very high relative to the average person's ability. Their insecurity arises because the high ability is paired with even higher self-esteem.

6.5 Aggression

Our final application, aggression, has received the lion's share of attention in the psychology literature dealing with insecurity. In a book titled *Evil: Inside Human Violence and Cruelty*, Baumeister (1997) argues that much if not most aggression results from people who "think well of themselves most of the time but who are vulnerable to frequent or large fluctuations in their self-esteem." For these people, Baumeister continues, "that miserable sinking feeling that goes with a drop in self-esteem is all too familiar, and they are on guard to avoid it. Even a slight hint or mild implication that questions their personal worth may elicit a strong response."

Our model formalizes aggression as a type of self-esteem-influencing choice. Suppose that the agent's partner acts in a way that he perceives as potentially lowering how he feels about himself. This threat could come about for multiple reasons. First, it can be akin to a low ϵ in Section 5.1, such as when the agent realizes that his partner is about to provide an unfavorable opinion or focus his attention on failures in his evidentiary base. The agent then aggresses — chooses a high e — which leads his partner to stop the behavior in question. This eliminates the looming negative light, protecting his feelings about himself. Second, the ego threat can be akin to a mood shock in Section 5.2, such as when the agent's partner is about to reveal a piece of information or otherwise get him to reevaluate himself. Here, the agent aggresses because he does not want uncertainty in his self-view. Furthermore, our framework identifies a role for small ego threats or slights in aggression. Such a small shock is likely to move the agent's self-assessment down by a bit, moving him closer to the point where his self-esteem collapses. At this point, the danger of a collapse is especially large, prompting aggression. The more unrealistic is the agent's high self-esteem, the more fragile it is, and hence the smaller is the slight that triggers a response. We are not aware of a formal model that predicts these features of aggressive behavior.

Our predictions appear to describe the qualitative findings of many researchers quite accurately. Kernis et al. (1989), for example, had 45 undergraduates provide multiple self-reports of self-esteem over a period of 1 week. One week later, they completed a battery of measures of aggression. The main finding from the study was that levels of aggression were particularly high among those with high, but also highly fluctuating, levels of self esteem. Berkowitz (1978) interviewed 65 incarcerated males who had been found guilty of inflicting serious bodily harm, but not in the course of a robbery. The purpose of the study was to test whether these men committed the acts of aggression for social rewards, or to live up to social-group expectations. The study failed to find support for this idea, but instead concluded that the men were "quick to see themselves as challenged and frequently interpreted someone else's remarks as belittling them," which "often infuriated" them and "stimulated them to lash out impulsively."

More quantitative evidence also supports our account. Again, narcissism (an indicator for high fragile self-esteem) is generally seen as associated with anger and aggression (e.g., Kernis, 2001). And while this topic is difficult to investigate experimentally, milder forms of aggression have been studied in economics experiments. Sebald and Walzl (2014) find that experimental subjects impose costly punishments for financially irrelevant subjective performance evaluations that they consider too low. Furthermore, an experimental subject's willingness to aggress in this way depends not only on his belief about his performance, but also on whether the belief is overconfident (Bellemare and Sebald, 2019). This is exactly in line with our prediction that given the agent's self-view, the more unrealistic this view is, the more prone he is to self-esteem-maintaining behaviors.

7 Concluding Remarks

Psychologists have documented the existence and consequences of fragility and insecurity, but have not discussed exactly what it means for self-esteem to be fragile. Taken at face value, the concept of fragility might seem to imply merely that starting from its current level, self-esteem could go up or down. The more common and consequential pattern of high-but-fragile self-esteem, however, represents a baseline level of self-esteem that can only go down. In this paper, we show that integrating the psychological concept of mood-congruent memory and the recent economic notion of personal equilibrium can help to make sense of this situation, explain a variety of findings from psychology, and shed new light on a range of phenomena of interest to economists.

To focus on the basic implications of fragile self-esteem, we have abstracted from several interesting issues. We have assumed that k is exogenous, but individuals may (consciously or subconsciously) choose the importance they attach to success. Claude Steele, one of the foremost researchers on education-related topics, provides an important example of choosing a low k in a thoughtful essay on “Race and the Schooling of Black Americans.” He writes about a process he calls “disidentification,” in which a black student, “at precisely the time when he would need to see school as a viable source of self-esteem resists measuring himself against its values and goals. He languishes there [i.e., in school], held by the law, perhaps even by his parents, but not allowing achievement to affect his view of himself.”

We have also assumed that self-esteem can be represented by a single scalar, but it would be more realistic to recognize that self-esteem is typically multidimensional. The same individual might, for example, experience different levels of self-esteem for different aspects of life such as work, family, friends, and athletics (Pelham, 1995), with different levels of importance. Furthermore, the implications for mood and memory are two-fold. On the one hand, self-esteem and the corresponding feelings about oneself in a particular dimension affect memories related to that dimension more than other memories. To the extent that this is the case, our single-dimensional model operates separately dimension by dimension. On the other hand, there is also an overall level of self-esteem and mood that is affected by, and affects, all individual dimensions of self-esteem. An obvious implication is that there are externalities across dimensions: a bad day at the office may affect one’s feelings about the family, and an ego-boosting hobby may help at work. Another notable implication, consistent with empirical research in psychology (Linville, 1985, 1987), is that people with more “diversified” self-esteem — one that is based on more dimensions of similar importance — will experience smaller swings of mood and overall self-esteem. Our perspective suggests that such diversification reduces mood volatility not only through the usual mechanism, but also by reducing the fragility of overall self-esteem as well as of each dimension of self-esteem. But our model also says that self-esteem in one domain can spill over and affect self-esteem in others, which can reduce the benefits of diversification.

Similarly, our two-outcome framework precludes us from talking about some interesting issues that arise when success and failure come in degrees. Even given the agent’s ability, his self-esteem is bound to be more fragile if his experiences have been more volatile, so that the experience the agent recalls has a great effect on mood. Consistent with this perspective, post-traumatic stress disorder can occur after a single terrible episode, but is far less likely to result from many minor failures.

Finally, we have not incorporated other, by now commonplace, behavioral elements into the model. We conjecture, in particular, that reference dependence generates qualitatively important novel predictions. Motivated by Kőszegi and Rabin (2009), suppose that feelings about oneself, and therefore memories or their interpretation, are sensitive not only to the level of, but also to recent changes in, self-esteem, with drops in self-esteem being particularly painful. This does not affect the long-run stable levels of self-esteem, when recent changes have not occurred. Crucially, however, such reference dependence renders either SEP, including the low one, vulnerable to *temporary* further collapse in self-esteem. If the agent's mood or self-esteem drops by a bit, he feels much worse, invoking the tailspin mechanism of our model and inducing a collapse. As he gets used to lower self-esteem, he feels less bad, gradually returning to the low SEP. The constant shadow of such a painful further collapse might make it especially difficult to get out of a low SEP, which, in turn, makes it all the more imperative to avoid getting to the low SEP in the first place.

References

Aditomo, Anindito, "Students' Response to Academic Setback:" Growth Mindset" as a Buffer against Demotivation.," *International Journal of Educational Psychology*, 2015, *4* (2), 198–222.

Alan, Sule, Teodora Boneva, and Seda Ertac, "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit," *Quarterly Journal of Economics*, 2019, *134* (3), 1121–1162.

Ames, Daniel R., Paul Rose, and Cameron P. Anderson, "The NPI-16 as a Short Measure of Narcissism," *Journal of Research in Personality*, 2006, *40* (4), 440–450.

Andreassen, Cecilie Schou, Holger Ursin, Hege R. Eriksen, and Ståle Pallesen, "The Relationship of Narcissism with Workaholism, Work Engagement, and Professional Position," *Social Behavior and Personality: An International Journal*, 2012, *40* (6), 881–890.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *Review of Economic Studies*, 2009, *76* (2), 451–469.

Avitabile, Ciro and Rafael De Hoyos, "The Heterogeneous Effect of Information on Student Performance: Evidence From a Randomized Control Trial in Mexico," *Journal of Development Economics*, 2018, *135*, 318–348.

Baumeister, Roy F., "Choking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance.," *Journal of Personality and Social Psychology*, 1984, *46* (3), 610.

_ , *Evil: Inside Human Violence and Cruelty*, W.H. Freeman, 1997.

_ **and Carolin J. Showers**, "A Review of Paradoxical Performance Effects: Choking Under Pressure in Sports and Mental Tests," *European Journal of Social Psychology*, 1986, *16* (4), 361–383.

Becker, Gary S. and Kevin M. Murphy, "A Theory of Rational Addiction," *Journal of Political Economy*, 1988, *96* (4), 675–700.

Bell, David E., "Disappointment in Decision Making Under Uncertainty," *Operations Research*, 1985, *33* (1), 1–27.

Bellemare, Charles and Alexander Sebald, "Self-Confidence and Reactions to Subjective Performance Evaluations," 2019. Working Paper.

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

_ **and** _ , "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

_ **and** _ , "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ **and** _ , "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 2016, *30* (3), 141–164.

Benoît, Jean-Pierre, Juan Dubra, and Don A. Moore, "Does the Better-Than-Average Effect Show That People Are Overconfident? Two Experiments.," *Journal of the European Economic Association*, 2015, *13* (2), 293–329.

Berglas, Steven and Edward E. Jones, "Drug Choice as a Self-Handicapping Strategy in Response to Noncontingent Success.," *Journal of Personality and Social Psychology*, 1978, *36* (4), 405.

Berkowitz, Leonard, "Is Criminal Violence Normative Behavior?: Hostile and Instrumental Aggression in Violent Incidents," *Journal of Research in Crime and Delinquency*, 1978, *15* (2), 148–161.

Bernheim, B. Douglas and Raphael Thomadsen, "Memory and Anticipation," *The Economic Journal*, 2005, *115* (503), 271–304.

Bettinger, Eric P., "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores," *Review of Economics and Statistics*, 2012, *94* (3), 686–698.

Bjørnstad, Roger, "Learned Helplessness, Discouraged Workers, and Multiple Unemployment Equilibria," *Journal of Socio-Economics*, 2006, *35* (3), 458–475.

Blanchard, Edward B., Jacqueline Jones-Alexander, Todd C. Buckley, and Catherine A. Forneris, "Psychometric Properties of the PTSD Checklist (PCL)," *Behaviour Research and Therapy*, 1996, *34* (8), 669–673.

Blaney, Paul H., "Affect and Memory: A Review," *Psychological Bulletin*, 1986, *99* (2), 229.

Bodner, Ronit and Drazen Prelec, "The Emergence of Private Rules in a Self-Signaling Model," *International Journal of Psychology*, 1996, *31*, 3652–3653.

Bodoh-Creed, Aaron L., "Mood, Memory, and the Evaluation of Asset Prices," *Review of Finance*, 2019, *24* (1), 227–262.

Bonanno, George A., Sandro Galea, Angela Bucciarelli, and David Vlahov, "What Predicts Psychological Resilience After Disaster? The Role of Demographics, Resources, and Life Stress.," *Journal of Consulting and Clinical Psychology*, 2007, *75* (5), 671.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and Representativeness," *Psychological Review*, 2021, *128* (1), 71–85.

_ , Nicola Gennaioli, and Andrei Shleifer, "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

Bower, Gordon H., "Mood and memory.," *American psychologist*, 1981, *36* (2), 129.

Brewin, Chris R., Bernice Andrews, and John D. Valentine, "Meta-Analysis of Risk Factors for Posttraumatic Stress Disorder in Trauma-Exposed Adults.," *Journal of Consulting and Clinical Psychology*, 2000, *68* (5), 748.

Buckley, Todd C., Edward B. Blanchard, and W. Trammell Neill, "Information Processing and PTSD: A Review of the Empirical Literature," *Clinical Psychology Review*, 2000, *20* (8), 1041–1065.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini, "Overconfidence and Social Signalling," *The Review of Economic Studies*, 2013, *80* (3), 949–983.

Bushong, Benjamin and Tristan Gagnon-Bartsch, "Learning with Misattribution of Reference Dependence," 2021. Working Paper.

Campbell, W. Keith, Adam S. Goodie, and Joshua D. Foster, "Narcissism, Confidence, and Risk Attitude," *Journal of Behavioral Decision Making*, 2004, *17* (4), 297–311.

Caro, Robert A., *The Path to Power: The Years of Lyndon Johnson I*, Vol. 1, Vintage, 2011.

Carrillo, Juan and Thomas Mariotti, "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies*, 2000, *67*, 529–544.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven, "Self-Confidence and Strategic Behavior," *Experimental Economics*, 2018, *21* (1), 72–98.

Chernow, Ron, *Alexander Hamilton*, Head of Zeus Ltd, 2016.

Clark, Andrew E. and Andrew J. Oswald, "Unhappiness and Unemployment," *The Economic Journal*, 1994, *104* (424), 648–659.

Clark, Malissa A., "Workaholism: It's Not Just Long Hours on the Job," *Psychological Science Agenda*, 2016.

_ , Ariel M. Lelchook, and Marcie L. Taylor, "Beyond the Big Five: How Narcissism, Perfectionism, and Dispositional Affect Relate to Workaholism," *Personality and Individual Differences*, 2010, *48* (7), 786–791.

_ , Jesse S. Michel, Ludmila Zhdanova, Shuang Y. Pui, and Boris B. Baltes, "All Work and No Play? A Meta-Analytic Examination of the Correlates and Outcomes of Workaholism," *Journal of Management*, 2016, *42* (7), 1836–1873.

Compte, Olivier and Andrew Postlewaite, "Confidence-Enhanced Performance," *American Economic Review*, December 2004, *94* (5), 1536–1557.

de la Rosa, Leonidas Enrique, "Overconfidence and Moral Hazard," *Games and Economic Behavior*, 2011, *73* (2), 429–451.

DellaVigna, Stefano and M. Daniele Paserman, "Job Search and Impatience," *Journal of Labor Economics*, 2005, *23* (3), 527–588.

Direnfeld, David M. and John E. Roberts, "Mood Congruent Memory in Dysphoria: The Roles of State Affect and Cognitive Style," *Behaviour Research and Therapy*, 2006, *44* (9), 1275–1285.

Eliaz, Kfir and Ran Spiegler, "Contracting with Diversely Naive Agents," *Review of Economic Studies*, 2006, *73* (3), 689–714.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann, "Associative Memory and Belief Formation," 2019. Working Paper.

Everaert, Jonas, Wouter Duyck, and Ernst H. W. Koster, "Attention, Interpretation, and Memory Biases in Subclinical Depression: A Proof-Of-Principle Test of the Combined Cognitive Biases Hypothesis," *Emotion*, 2014, *14* (2), 331–340.

Farnham, Shelly D., Anthony G. Greenwald, and Mahzarin R. Banaji, "Implicit Self-Esteem," in D. Abrams M.A. Hogg, ed., *Social Identity and Social Cognition*, Blackwell Publishing, 1999, pp. 230–248.

Feather, Norman T., *The Psychological Impact of Unemployment*, Springer Science & Business Media, 2012.

Ferrari, Joseph R. and Ted Thompson, "Impostor Fears: Links with Self-Presentational Concerns and Self-Handicapping Behaviours," *Personality and Individual Differences*, 2006, *40* (2), 341–352.

Forgas, Joseph P., "Mood Effects on Cognition: Affective Influences on the Content And Process of Information Processing and Behavior," in "Emotions and Affect in Human Factors and Human-Computer Interaction," Elsevier, 2017, chapter 3, pp. 89–122.

_ **and Gordon H. Bower**, "Mood Effects on Person-Perception Judgments," *Journal of Personality and Social Psychology*, 1987, *53* (1), 53–60.

Fryer Jr., Roland G., "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *The Quarterly Journal of Economics*, 2011, *126* (4), 1755–1798.

_ , "Information and Student Achievement: Evidence From a Cellular Phone Experiment," Technical Report, National Bureau of Economic Research 2013.

Gervais, Simon and Terrance Odean, "Learning To Be Overconfident," *Review of Financial Studies*, 2001, *14* (1), 1–27.

Gilboa-Schechtman, Eva, Dana Erhard-Weiss, and Pablo Jeczemien, "Interpersonal Deficits Meet Cognitive Biases: Memory for Facial Expressions in Depressed and Anxious Men and Women," *Psychiatry Research*, 2002, *113* (3), 279–293.

Goldsmith, Arthur H., Jonathan R. Veum, and William Darity Jr., "The Psychological Impact of Unemployment and Joblessness," *Journal of Socio-Economics*, 1996, *25* (3), 333–358.

Gotlib, Ian H. and Jutta Joormann, "Cognition and Depression: Current Status and Future Directions," *Annual Review of Clinical Psychology*, 2010, *6*, 285–312.

Greenberg, Jerald, "Unattainable Goal Choice as a Self-Handicapping Strategy 1," *Journal of Applied Social Psychology*, 1985, *15* (2), 140–152.

Grossman, Zachary and Joël J. van der Weele, "Self-Image and Willful Ignorance in Social Decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 2018, *86* (4), 1159–1214.

Heimpel, Sara A., Joanne V. Wood, Margaret A. Marshall, and Jonathon D. Brown, "Do People with Low Self-Esteem Really Want to Feel Better? Self-Esteem Differences in Motivation to Repair Negative Moods," *Journal of Personality and Social Psychology*, 2002, *82* (1), 128–147.

Helliwell, John F. and Haifang Huang, "New Measures of the Costs of Unemployment: Evidence from the Subjective Well-Being of 3.3 million Americans," *Economic Inquiry*, 2014, *52* (4), 1485–1502.

Hetts, John J., Michiko Sakuma, and Brett W. Pelham, "Two Roads to Positive Regard: Implicit and Explicit Self-Evaluation and Culture," *Journal of Experimental Social Psychology*, 1999, *35* (6), 512–559.

Higgins, Raymond L., Charles Richard Snyder, and Steven Berglas, *Self-Handicapping: The Paradox that Isn't*, Springer Science & Business Media, 2013.

Hoffman, Mitchell and Stephen V. Burks, "Worker Overconfidence: Field Evidence and Implications for Employee Turnover and Firm Profits," *Quantitative Economics*, 2020, *11* (1), 315–348.

Huffman, David, Collin Raymond, and Julia Shvets, "Persistent Overconfidence and Biased Memory: Evidence from Managers," 2019. Working Paper.

Isaacson, Walter, *Kissinger: A biography*, Simon and Schuster, 2005.

Isen, Alice M, Thomas E Shalker, Margaret Clark, and Lynn Karp, "Affect, Accessibility of Material in Memory, and Behavior: A Cognitive Loop?," *Journal of Personality and Social Psychology*, 1978, *36* (1), 1–12.

James, William, *The Principles of Psychology*, Vol. 1, Holt, 1890.

Jensen, Robert, "The (Perceived) Returns to Education and the Demand for Schooling," *The Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

Johnson, Dominic D.P. and James H. Fowler, "The Evolution of Overconfidence," *Nature*, 2011, *477* (7364), 317.

_ , Nils B. Weidmann, and Lars-Erik Cederman, "Fortune Favours the Bold: An Agent-Based Model Reveals Adaptive Advantages of Overconfidence in War," *PLOS One*, 2011, *6* (6), e20851.

Johnson, Eric J. and Amos Tversky, "Affect, Generalization, and the Perception of Risk," *Journal of Personality and Social Psychology*, 1983, *45*, 20–31.

Jones, Del, "Could Insecurity Be the Secret to CEOs' Success?," *USA Today*, 2007, pp. 1–2.

Josephson, Braden R., "Mood Regulation and Memory: Repairing Sad Moods with Happy Memories," *Cognition & Emotion*, 1996, *10* (4), 437–444.

Kernis, Michael H., "Following the Trail from Narcissism to Fragile Self-Esteem," *Psychological Inquiry*, 2001, *12* (4), 223–225.

_ , Bruce D. Grannemann, and Lynda C. Barclay, "Stability and Level of Self-Esteem as Predictors of Anger Arousal and Hostility," *Journal of Personality and Social Psychology*, 1989, *56* (6), 1013–1022.

_ , Teresa A. Abend, Brian M. Goldman, Ilan Shrira, Andrew N. Paradise, and Christian Hampton, "Self-Serving Responses Arising from Discrepancies Between Explicit and Implicit Self-Esteem," *Self and Identity*, 2005, *4* (4), 311–330.

Kőszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 2006, *4* (4), 673–707.

_ , "Utility from Anticipation and Personal Equilibrium," *Economic Theory*, 2010, *44* (3), 415–444.

_ **and Matthew Rabin**, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 2006, *121* (4), 1133–1166.

_ **and** _ , "Reference-Dependent Risk Attitudes," *American Economic Review*, 2007, *97* (4), 1047–1073.

_ **and** _ , "Reference-Dependent Consumption Plans," *American Economic Review*, 2009, *99* (3), 909–936.

Koole, Sander L., Ap Dijksterhuis, and Ad Van Knippenberg, "What's in a Name: Implicit Self-Esteem and the Automatic Self.," *Journal of Personality and Social Psychology*, 2001, *80* (4), 669.

Kőszegi, Botond and Wei Li, "Drive and Talent," *Journal of the European Economic Association*, 2008, *6* (1), 210–236.

Krueger, Alan B. and Andreas I. Mueller, "Job Search and Unemployment Insurance: New Evidence from Time Use Data," *Journal of Public Economics*, 2010, *94* (3-4), 298–307.

_ **and** _ , "The Lot of the Unemployed: A Time Use Perspective," *Journal of the European Economic Association*, 2012, *10* (4), 765–794.

Laibson, David I., "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, *112* (2), 443–477.

Lamba, Shakti and Vivek Nityananda, "Self-Deceived Individuals are Better at Deceiving Others," *PloS one*, 2014, *9* (8), e104562.

Langford, Joe and Pauline Rose Clance, "The Imposter Phenomenon: Recent Research Findings Regarding Dynamics, Personality and Family Patterns and their Implications for Treatment.," *Psychotherapy: Theory, Research, Practice, Training*, 1993, *30* (3), 495.

Lavecchia, Adam M., Heidi Liu, and Philip Oreopoulos, "Behavioral Economics of Education: Progress and Possibilities," in "Handbook of the Economics of Education," Vol. 5, Elsevier, 2016, pp. 1–74.

Lazarus, Richard S. and Susan Folkman, "Coping and Adaptation," *The Handbook of Behavioral Medicine*, 1984, *282325.*

Leuven, Edwin, Hessel Oosterbeek, and Bas Van der Klaauw, "The Effect of Financial Rewards on Students' Acheivements: Evidence From a Randomized Experiment," *Journal of the European Economic Association*, 2010, *8* (6), 1243–1265.

Linville, Patricia W., "Self-Complexity and Affective Extremity: Don't Put All of Your Eggs in One Cognitive Basket," *Social Cognition*, 1985, *3* (1), 94–120.

_ , "Self-Complexity as a Cognitive Buffer Against Stress-Related Illness and Depression," *Journal of Personality and Social Psychology*, 1987, *52* (4), 663–676.

Loomes, Graham and Robert Sugden, "Disappointment and Dynamic Consistency in Choice under Uncertainty," *Review of Economic Studies*, 1986, *53* (2), 271–282.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *Journal of Finance*, 2005, *60* (6), 2661–2700.

Maslow, Abraham Harold, "The Dynamics of Psychological Security-Insecurity.," *Character & Personality; A Quarterly for Psychodiagnostic & Allied Studies*, 1942.

Matt, Georg E., Carmelo Vázquez, and W. Keith Campbell, "Mood-Congruent Recall of Affectively Toned Stimuli: A Meta-Analytic Review," *Clinical Psychology Review*, 1992, *12* (2), 227–255.

Mayer, John D., Laura J. McCormick, and Sara E. Strong, "Mood-Congruent Memory and Natural Mood: New Evidence," *Personality and Social Psychology Bulletin*, 1995, *21* (7), 736–746.

_ , Yvonne N. Gaschke, Debra L. Braverman, and Temperance W Evans, "Mood-Congruent Judgment is a General Effect.," *Journal of Personality and Social Psychology*, 1992, *63* (1), 119.

McGuigan, Martin, Sandra McNally, and Gill Wyness, "Student Awareness of Costs and Benefits of Educational Decisions: Effects of an Information Campaign," *Journal of Human Capital*, 2016, *10* (4), 482–519.

Miranda, Regina and John Kihlstrom, "Mood Congruence in Childhood and Recent Autobiographical Memory," *Cognition & Emotion*, 2005, *19* (7), 981–998.

Morf, Carolyn C. and Frederick Rhodewalt, "Unraveling the Paradoxes of Narcissism: A Dynamic Self-Regulatory Processing Model," *Psychological Inquiry*, 2001, *12* (4), 177–196.

Mullainathan, Sendhil, "Thinking Through Categories," 2000. Working Paper.

_ , "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2002, *117* (3), 735–774.

_ , Joshua Schwartzstein, and Andrei Shleifer, "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 2008, *123* (2), 577–619.

O'Donoghue, Ted and Matthew Rabin, "Doing It Now or Later," *American Economic Review*, 1999, *89* (1), 103–124.

Orphanides, Athanasios and David Zervos, "Rational Addiction With Learning and Regret," *Journal of Political Economy*, 1995, *103* (4), 739–758.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey, "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease," *American Economic Review*, 2013, *103* (2), 804–30.

Pelham, Brett W., "Self-Investment and Self-Esteem: Evidence for a Jamesian model of Self-Worth," *Journal of Personality and Social Psychology*, 1995, *69* (6), 1141.

Prelec, Drazen and Ronit Bodner, "Self-Signaling and Self-Control," in "Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice," New York, NY, US: Russell Sage Foundation, 2003, pp. 277–298.

Raskin, Robert and Howard Terry, "A Principal-Components Analysis of the Narcissistic Personality Inventory and Further Evidence of its Construct Validity," *Journal of Personality and Social Psychology*, 1988, *54* (5), 890–902.

Sakulku, Jaruwan, "The Impostor Phenomenon," *The Journal of Behavioral Science*, 2011, *6* (1), 75–97.

Santos-Pinto, Luis and Joel Sobel, "A Model of Positive Self-Image in Subjective Assessments," *American Economic Review*, 2005, *95* (5), 1386–1402.

Schelling, Thomas C., "The Mind as a Consuming Organ," *The Multiple Self*, 1987, pp. 177–96.

Sebald, Alexander and Markus Walzl, "Subjective Performance Evaluations and Reciprocity in Principal-Agent Relations," *Scandinavian Journal of Economics*, 2014, *116* (2), 570–590.

Shui, Haiyan and Lawrence M. Ausubel, "Time Inconsistency in the Credit Card Market," 2005. Working Paper.

Smith, Adam, *The Theory of Moral Sentiments*, London: A. Millar, 1759.

Spiegler, Ran, "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

Spinnewijn, Johannes, "Insurance and Perceptions: How to Screen Optimists and Pessimists," *Economic Journal*, 2013, *123* (569), 606–633.

_ , "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs," *Journal of the European Economic Association*, 2015, *13* (1), 130–167.

Turner, R. Jay, Blair Wheaton, and Donald A. Lloyd, "The Epidemiology of Social Stress," *American Sociological Review*, 1995, pp. 104–125.

Uluduz, Hatice and Ilhan Gunbayi, "Growth Mindset in the Classroom," *European Journal of Education Studies*, 2018.

Wachter, Jessica A. and Michael J. Kahana, "A Retrieved-Context Theory of Financial Decisions," 2021. Working Paper.

Watkins, Philip C., Andrew Mathews, Donald A. Williamson, and Richard D. Fuller, "Mood-Congruent Memory in Depression: Emotional Priming or Elaboration?," *Journal of Abnormal Psychology*, 1992, *101* (3), 581.

_ , Karen Vache, Steven P. Verney, and Andrew Mathews, "Unconscious Mood-Congruent Memory Bias in Depression.," *Journal of Abnormal Psychology*, 1996, *105* (1), 34.

Winkelmann, Liliana and Rainer Winkelmann, "Why are the Unemployed so Unhappy? Evidence from Panel Data," *Economica*, 1998, *65* (257), 1–15.

Young, Valerie, *The Secret Thoughts of Successful Women: Why Capable People Suffer from the Impostor Syndrome and How to Thrive in Spite of it*, Crown Pub, 2011.

Zábojnik, Ján, "A Model of Rational Bias in Self-Assessments," *Economic Theory*, 2004, *23* (2), 259–282.

Appendices

A A Model of Building Self-Esteem

A.1 Formal Model

This section discusses the possibility that the agent's ability can be developed through effort. For instance, a child attempting basketball may realize that his shooting and dribbling improve quickly with practice. It is also realistic to assume that the child's self-esteem is based not purely on his innate talent for basketball, but on a projection of how skilled he is likely to become with practice. This possibility introduces a scope for building up one's self-esteem — but we show that such a process is fraught with fragility.

We normalize $k = 1$, and suppose that the agent can exert effort $e \in \{0, 1\}$ to improve his ability, with a_0 and a_1 denoting his levels of ability given $e = 0$ and $e = 1$, respectively, and $\delta = a_1 - a_0 > 0$ denoting the contribution of effort to ability. To focus on the most interesting set of possibilities, we assume that whether or not he exerts effort, the agent has fragile self-esteem with SEPs $\tilde{a}_e^{low}$ and $\tilde{a}_e^{high}$, and we denote the unstable fixed point by $\tilde{a}_e^u$. As in Section 5, the agent's effort level determines the distribution of shocks for SEP determination. To make our points most clearly, we make two simplifying assumptions. First, the agent is not subject to any mood shock if he chooses $e = 0$, and he experiences a mood shock $-\epsilon < 0$ — i.e., a challenge to his self-view — if he chooses $e = 1$. Second, when starting from a SEP, the agent finds it optimal to exert effort if and only if the SEP is high and stable to the negative shock ϵ. That the agent is more likely to exert effort in a high SEP is consistent with the setup of Section 4 that his cost of effort is lower when he feels good about himself, or his perceived return to effort is higher when he has a higher opinion of himself. That he is not willing to exert effort when the shock would destroy his high SEP is consistent with the self-esteem-influencing motive, as discussed in Section 5. An alternative interpretation is that, if the agent exerts effort in a high SEP that is not stable to the shock, then he destroys his SEP and hence cannot persistently be in it. While simplistic, this formulation allows us to isolate the role of emotional fragility in building self-esteem by ignoring any long-run human-capital-building motive to exert effort. Including these motives in our model would not affect any of our qualitative conclusions.

Because effort can impact ability and therefore SEP determination, and conversely SEP determination can impact effort, the appropriate concept of personal equilibrium is more complex than in our previous models. We define an e-SEP as a pair $(e, \tilde{a})$ such that (1) given that the proportion of successes in the evidentiary base is a_e, $\tilde{a}$ is a SEP; (2) given $\tilde{a}$, e is optimal. The most straightforward interpretation is that e is a steady-state level of effort in a dynamic setting. If the agent regularly exerts effort e, then his ability must be a_e and his evidentiary base must be filled with memories corresponding to this ability. Hence, he must be in a SEP consistent with a true ability of a_e. Furthermore, for him to be willing to consistently exert effort e, this effort must be optimal given the SEP he is in.

An alternative interpretation, especially appropriate for a young person such as a school-age child, is forward-looking. In this interpretation, e is the effort level the agent expects to exert on a regular basis. Since he does not have many relevant memories himself, his evidentiary base for what will happen is based on his beliefs, constructed for instance from his observations of the experiences of others, about how someone regularly exerting effort e fares. As above, his SEP must then be based on this evidentiary base, and our definition of e-SEP says that exerting effort e must be optimal given this evidentiary base and selected SEP.

To illustrate both the definition and the use of the concept of e-SEP, we return to the basketball hopeful above. The child can practice a little or a lot, and practicing a lot can generate emotional drops when things do not go well. Furthermore, in a low SEP, the child does not practice, either because he perceives the return to be low, or because feeling bad about himself is de-motivating (or both). Hence, one e-SEP involves low self-esteem and no practice. More interestingly, there may be two e-SEPs with high self-esteem. If the child believes that he will practice a lot, then his high SEP is very stable — robust to a negative shock — and he will indeed be willing to practice. Intuitively, because he entered the state assured by the fact that he will eventually get good, a shock does not destroy his self-esteem. If he believes that he will not practice, in contrast, then he starts off at a lower self-esteem, so that a high SEP is more unstable, and practicing or otherwise exposing himself to challenges is risky. As a result, he does not practice. Although he ends up at a high SEP, he remains a mediocre player.

Proposition 8 formally summarizes these results:

Proposition 8. *i) an e-SEP* $(0, \tilde{a}_0^{low})$ *always exists; ii) an e-SEP* $(1, \tilde{a}_1^{high})$ *exists if and only if* $\tilde{a}_1^{high} - \tilde{a}_1^u \geq \epsilon$*; iii) an e-SEP* $(0, \tilde{a}_0^{high})$ *exists if and only if* $\tilde{a}_0^{high} - \tilde{a}_0^u < \epsilon$.

There are two mathematically straightforward, but economically important comparative statics to note. First, the greater is the threat ϵ, the less likely it is that an e-SEP $(1, \tilde{a}_1^{high})$ exists. Second, and more importantly, fixing a_0, the higher is the contribution of effort to ability, δ, the more likely it is that an e-SEP $(1, \tilde{a}_1^{high})$ exists. The higher is the amount by which effort raises ability, the more stable is the high SEP resulting from high effort, making the agent more willing to exert effort. In fact, in the forward-looking interpretation of our model, an increase in the agent's *belief* about δ has the same effect: by raising his expectation about what will happen, it improves his evidentiary base, raising his self-esteem and thereby his willingness to exert effort.

A.2 Growth versus Fixed Mindset

We can naturally define mindset as a student's belief about δ: students who believe that δ is low have a fixed mindset, and students who believe that δ is high have a growth mindset. Then, our model says that an e-SEP in which the student studies hard is more likely to exist for those with a growth mindset. Furthermore, consistent with the evidence discussed in Section 6.2 , the difference between the two mindsets arises exactly when confronted with failure: studying is robust to such an experience for a student with the growth mindset, but not for a student with a fixed mindset. In our setting, therefore, an intervention that instills a growth mindset involves *both* inducing a belief that δ is high and setting up an e-SEP in which the student has high and relatively secure self-esteem.

B Proofs

Proof of Fact 1.

Because $g(\cdot)$ is symmetric around zero, $g(0-k\tilde{a}) = g(k\tilde{a})$, and hence $E[s|k\tilde{a}] = \frac{g(k-k\tilde{a})ka}{g(k-k\tilde{a})a+g(0-k\tilde{a})(1-a)} = \frac{g(k-k\tilde{a})ka}{g(k-k\tilde{a})a+g(k\tilde{a})(1-a)}$. Note that for any binary database, each fixed database is represented by a single parameter $a \in [0,1]$. For notational convenience, let $G_a(\tilde{a}) \equiv \text{Prob}[\text{success}|\tilde{a}] = E[s|k\tilde{a}]/k$ for each fixed database with a; $\tilde{a}$ is a SEP if it is a locally stable solution to $G_a(\tilde{a}) = \tilde{a}$.

We first show that $G_a(\tilde{a})$ is weakly increasing in $\tilde{a}$. By taking the derivative of $G_a(\tilde{a})$ with respect to $\tilde{a}$,

$$\begin{aligned} G'_a(\tilde{a}) &= \frac{-kg'(k-k\tilde{a})a[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)] - g(k(1-\tilde{a}))a[-kg'(k-k\tilde{a})a+kg'(k\tilde{a})(1-a)]}{[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]^2} \\ &= \frac{-ka(1-a)[g'(k\tilde{a})g(k-k\tilde{a})+g'(k-k\tilde{a})g(k\tilde{a})]}{[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]^2}. \end{aligned}$$

As $g'(x) \leq 0$ for $x > 0$ by the assumption on $g(\cdot)$, $G'_a(\tilde{a}) \geq 0$.

Now we prove Fact 1. Because $G_0(\tilde{a}) = 0$ (respectively $G_1(\tilde{a}) = 1$) for any $k > 0$, $g(\cdot)$ and $\tilde{a} \in [0,1]$, if $a = 0$ (respectively $a = 1$), then the solution is globally stable and equals 0 (respectively 1). Suppose now $a \in (0,1)$. Because $G_a(0) > 0$, $G_a(1) < 1$, and $G_a(\tilde{a})$ is continuous and increasing in $\tilde{a}$, there exists $\tilde{a}^* \in (0,1)$ such that $G_a(\tilde{a}^*) = \tilde{a}^*$ with $G'_a(\tilde{a}^*) \leq 1$. Because of the assumption that any solution to the equation $G_a(\tilde{a}) = \tilde{a}$ is locally unique, $G'_a(\tilde{a}^*) \neq 1$ and hence $G'_a(\tilde{a}^*) < 1$, implying that $\tilde{a}^*$ is a locally-stable solution to the equation. □

Proof of Proposition 1.

As derived in Fact 1, $G_a(\tilde{a}) = \frac{g(k-k\tilde{a})a}{g(k-k\tilde{a})a+g(k\tilde{a})(1-a)}$. We show that $G_a(\tilde{a})$ has at most one inflection point, which ensures that there are at most two SEPs.

As shown in Fact 1, $G'_a(\tilde{a}) \geq 0$. Also,

$$\begin{aligned} G''_a(\tilde{a}) =& \frac{k^2a(1-a)}{[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]^3} \\ &\cdot \big\{2[g'(k\tilde{a})g(k-k\tilde{a})+g'(k-k\tilde{a})g(k\tilde{a})]\cdot[-g'(k-k\tilde{a})a+g'(k\tilde{a})(1-a)] \\ &\quad -[g''(k\tilde{a})g(k-k\tilde{a})-g''(k-k\tilde{a})g(k\tilde{a})]\cdot[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]\big\}. \end{aligned}$$

Note that the term multiplied by the curly brackets is always positive. Hence, if a derivative of the terms in the curly brackets of $G''_a(\tilde{a})$ is always non-positive, then $G''_a(\tilde{a})$ crosses zero at most once, implying that $G_a(\tilde{a})$ has at most one inflection point. Note that the derivative is:

$$\begin{aligned}
&-k[g'''(k\tilde{a})g(k-k\tilde{a})+g'''(k-k\tilde{a})g(k\tilde{a})]\cdot[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]\\
&+3k[g'(k\tilde{a})g(k-k\tilde{a})+g'(k-k\tilde{a})g(k\tilde{a})]\cdot[g''(k-k\tilde{a})a+g''(k\tilde{a})(1-a)]. \qquad (1)
\end{aligned}$$

As $g(x) \geq 0$ and $g'(x) \leq 0$, (1) is non-positive when $g''(x) \geq g(x)(\geq 0)$ and $g'''(x) \geq 0$ for all $x > 0$. Now suppose that $g'''(x) < 0$ for some $x > 0$. Even in this case, if $g'(x) \leq g'''(x)(< 0)$ and $g''(x) \geq g(x)(\geq 0)$ hold for all $x > 0$, then (1) is non-positive. □

Proof of Proposition 2.

I. By definition, $G_0(\tilde{a}) = 0$ and $G_1(\tilde{a}) = 1$ for any $k > 0$ and $\tilde{a} \in [0, 1]$. Therefore, if $a = 0$ or $a = 1$, then the SEP is unique. □

II. Because $G'_a(\tilde{a}) = \frac{-ka(1-a)[g'(k\tilde{a})g(k-k\tilde{a})+g'(k-k\tilde{a})g(k\tilde{a})]}{[g(k-k\tilde{a})a+g(k\tilde{a})(1-a)]^2}$, $G'_a(\tilde{a})$ approaches 0 as $k \to \infty$. Thus, there exists $\underline{k} > 0$ such that if $k < \underline{k}$, then $G'_a(\tilde{a}) < 1$ for all $\tilde{a} \in [0, 1]$. This ensures that the agent does not have multiple SEPs. □

III. Because of the assumption that a solution to the equation $\tilde{a} = G_a(\tilde{a})$ is always locally unique, it suffices to show that there exists an unstable fixed point $G_a(\tilde{a}) = \tilde{a} \in (0, 1)$ for any sufficiently large $k > 0$. Since $G_a(0) > 0$ and $G_a(1) < 1$ for any $a \in (0, 1)$, such an unstable fixed point $\tilde{a}$ exists if for any sufficiently large $k > 0$ there exist $\tilde{a}_L, \tilde{a}_H \in (0, 1)$ such that $\tilde{a}_L < \tilde{a}_H$, $G_a(\tilde{a}_L) < \tilde{a}_L$, and $G_a(\tilde{a}_H) > \tilde{a}_H$ (i.e., $G_a(\tilde{a})$ crosses the 45-degree line from below at $\tilde{a} \in (\tilde{a}_L, \tilde{a}_H)$).

For the later use, we prove a slightly general version for $\tilde{a}_L$: for any $\epsilon \in (0, 1]$, there is $\bar{k}_L > 0$ such that for any $k > \bar{k}_L$ a low SEP $\tilde{a}^{low}$ is smaller than ϵ. Take $\tilde{a}_L = \frac{\epsilon}{k}$ for any $k > 1$. Because $G_a(\tilde{a}_L) = \frac{g(k-k\tilde{a}_L)a}{g(k-k\tilde{a}_L)a+g(k\tilde{a}_L)(1-a)} = \frac{ag(k-\epsilon)}{ag(k-\epsilon)+g(\epsilon)(1-a)}$, we have $G_a(\tilde{a}_L) - \tilde{a}_L = \frac{ag(k-\epsilon)}{ag(k-\epsilon)+g(\epsilon)(1-a)} - \frac{\epsilon}{k} = \frac{a(k-\epsilon)g(k-\epsilon)-\epsilon g(\epsilon)(1-a)}{k[ag(k-\epsilon)+g(\epsilon)(1-a)]}$. Because the right hand side becomes negative as $k \to +\infty$, for a sufficiently large k, there exists $\tilde{a}_L = \frac{\epsilon}{k}$ such that $G_a(\tilde{a}_L) < \tilde{a}_L$.

Similarly, we prove a slightly general version for $\tilde{a}_H$: for any $\epsilon \in (0, 1]$, there is $\bar{k}_H > 0$ such that for any $k > \bar{k}_H$ a high SEP $\tilde{a}^{high}$ is larger than $\frac{k-\epsilon}{k}$. Take $\tilde{a}_H = \frac{k-\epsilon}{k}$ for any $k > 1$. Because $G_a(\tilde{a}_H) = \frac{g(k-k\tilde{a}_H)a}{g(k-k\tilde{a}_H)a+g(k\tilde{a}_H)(1-a)} = \frac{g(\epsilon)a}{g(\epsilon)a+g(k-\epsilon)(1-a)}$, we have $G_a(\tilde{a}_H) - \tilde{a}_H = \frac{g(\epsilon)a}{g(\epsilon)a+g(k-\epsilon)(1-a)} -$

$\frac{k-\epsilon}{k} = \frac{g(\epsilon)a\epsilon-(k-\epsilon)g(k-\epsilon)(1-a)}{k[g(\epsilon)a+g(k-\epsilon)(1-a)]}$. Because the right-hand side becomes positive as $k \to +\infty$, for a sufficiently large k, there exists $\tilde{a}_H = \frac{k-\epsilon}{k}$ such that $G_a(\tilde{a}_H) > \tilde{a}_H$.

Hence, for any $a \in (0,1)$, there is a $\overline{k} = \max\{\bar{k}_L, \bar{k}_H\}$ such that if $k > \overline{k}$, then the agent has fragile self-esteem. □

Proof of Proposition 3.

I. By Proposition 2 Part III, there are two SEPs in this case. Moreover, one of these SEPs is smaller than $\tilde{a}_L$ and another is larger than $\tilde{a}_H$, where $\tilde{a}_L$ and $\tilde{a}_H$ are specified in the proof of Proposition 2 Part III. Thus, if we take $k > \max\{a, \frac{1}{1-a}, \overline{k}\}$ where $\overline{k}$ is specified in Proposition 2 Part III, one SEP is strictly smaller than a while another SEP is strictly larger than a. □

II. It is straightforward to show that $\lim_{k\to 0} G_a(\tilde{a}) = a$. In the proof of Proposition 2 Part III with setting $\epsilon = 1$, we showed that for a sufficiently large k, there are two SEPs where one SEP is smaller than $\frac{1}{k}$ and another SEP is larger than $\frac{k-1}{k}$. Taking $k \to \infty$ completes the proof. □

III. In the proof of Proposition 2 Part III, we showed that for any $\epsilon \in (0,1]$, if $k > 1$ is sufficiently large, a high SEP $\tilde{a}^{high}$ is larger than $\frac{k-\epsilon}{k}$. Take $\bar{\epsilon} = a' - a > 0$. Because $\tilde{a}^{high\ \prime} < 1$ and $\tilde{a}^{high} > \frac{k-\bar{\epsilon}}{k}$ for any sufficiently large $k > 1$, we obtain $\tilde{a}^{high\ \prime} - \tilde{a}^{high} < 1 - \frac{k-\bar{\epsilon}}{k} = \frac{a'-a}{k} < a' - a$. Therefore, $\tilde{a}^{high} - a > \tilde{a}^{high\ \prime} - a'$. The result for the low SEP can be shown in the same manner. □

Proof of Proposition 4.

We first show that, if $|\tilde{a} - a| \geq 1/2$, then the agent has fragile self-esteem. Suppose otherwise: the agent has a unique SEP. By Proposition 2 Part II, $|\tilde{a} - a| \geq 1/2$ implies $a \in (0,1)$. Because $G_a(a) = \frac{g(k-ka)a}{g(k-ka)a+g(ka)(1-a)}$, for any $a \in (0,1)$, $G_a(a) > a$ if $a > 1/2$; $G_a(a) = a$ if $a = 1/2$; and $G_a(a) < a$ if $a < 1/2$. Because of the assumption that a solution to the equation $\tilde{a} = G_a(\tilde{a})$ is always locally unique, the agent's unique SEP is $\tilde{a} > 1/2$ if $a > 1/2$; $\tilde{a} = a$ if $a = 1/2$; and $\tilde{a} < 1/2$ if $a < 1/2$. In sum, we obtain $|\tilde{a} - a| < 1/2$ for any $a \in [0,1]$, a contradiction.

We now show that, if $\tilde{a} - a \leq -1/2$, then the agent is in a low fragile SEP. By Proposition 2 Part II and $\tilde{a} \in [0,1]$, it implies $a \in (1/2, 1)$. Because $G_a(a) > a$ if $a \in (1/2, 1)$ and the assumption that a solution to the equation $\tilde{a} = G_a(\tilde{a})$ is always locally unique, there is a SEP $\tilde{a}^{high} > a > 1/2$.

Therefore, if $\tilde{a} - a \leq -1/2$, then the agent must be in a low SEP. The case in which $\tilde{a} - a \geq 1/2$ can be shown in the same manner. $\square$

Proof of Proposition 5.

Suppose there are two SEPs. Because of the assumption that a solution to the equation $\tilde{a} = G_a(\tilde{a})$ is always locally unique, $G_a(0) \geq 0$, and $G_a(1) \leq 1$, there is one unstable fixed point $\tilde{a}^u = G_a(\tilde{a}^u) \in (0,1)$. Note that a small increase in a makes the high SEP more stable and the low SEP less stable if $\frac{d\tilde{a}^u}{da} < 0$.

Let denote $f(\tilde{a}^u, a) = G_a(\tilde{a}^u) - \tilde{a}^u = 0$. Note that we have $\frac{\partial f}{\partial \tilde{a}^u} = G'_a(\tilde{a}^u) - 1 > 0$. Hence, by the Implicit Function Theorem, we obtain

$$\begin{aligned}
\frac{d\tilde{a}^u}{da} &= -\frac{\partial f/\partial a}{\partial f/\partial \tilde{a}^u} \\
&= -\frac{1}{G'_a(\tilde{a}^u) - 1} \cdot \frac{g(k - k\tilde{a}^u)[g(k - k\tilde{a}^u)a + g(k\tilde{a}^u)(1-a)] - g(k - k\tilde{a}^u)a[g(k - k\tilde{a}^u) - g(k\tilde{a}^u)]}{[g(k - k\tilde{a}^u)a + g(k\tilde{a}^u)(1-a)]^2} \\
&= -\frac{1}{G'_a(\tilde{a}^u) - 1} \cdot \frac{g(k - k\tilde{a}^u)g(k\tilde{a}^u)}{[g(k - k\tilde{a}^u)a + g(k\tilde{a}^u)(1-a)]^2} < 0. \quad \square
\end{aligned}$$

Proof of Proposition 6.

I. Note that the agent's effort is characterized by $c'(e) = k\tilde{a}$. In what follows, we show that $k\frac{d\tilde{a}}{dk}$ converges to zero as $k \to 0$. Combined with Proposition 3 Part III, it implies that $\lim_{k\to 0} \frac{dk\tilde{a}}{dk} = \lim_{k\to 0} \tilde{a} + \lim_{k\to 0} k\frac{d\tilde{a}}{dk} = a > 0$, and hence, there exists a sufficiently small $\underline{k} > 0$ such that $\frac{dk\tilde{a}}{dk} > 0$ for any $k < \underline{k}$.

Denote by $f(\tilde{a}, k) = G_a(\tilde{a}) - \tilde{a} = 0$. Note that we have $\frac{\partial f}{\partial \tilde{a}} = G'_a(\tilde{a}) - 1 < 0$ by the definition of SEP. Hence, by the implicit function theorem, we obtain

$$\begin{aligned}
k\frac{d\tilde{a}}{dk} &= -k\frac{\partial f/\partial k}{\partial f/\partial \tilde{a}} \\
&= -k\frac{1}{G'_a(\tilde{a}) - 1} \cdot \frac{a(1-\tilde{a})g'(k - k\tilde{a})[g(k - k\tilde{a})a + g(k\tilde{a})(1-a)] - ag(k - k\tilde{a})[a(1-\tilde{a})g'(k - k\tilde{a}) + (1-a)\tilde{a}g(k\tilde{a})]}{[g(k - k\tilde{a})a + g(k\tilde{a})(1-a)]^2} \\
&= -k\frac{[g(k - k\tilde{a})a + g(k\tilde{a})(1-a)]^2}{-ka(1-a)[g'(k\tilde{a})g(k - k\tilde{a}) + g'(k - k\tilde{a})g(k\tilde{a})] - [g(k - k\tilde{a})a + g(k\tilde{a})(1-a)]^2} \\
&\quad \cdot \frac{a(1-a)[g'(k - k\tilde{a})g(k\tilde{a})(1-\tilde{a}) - g(k - k\tilde{a})g'(k\tilde{a})\tilde{a}]}{[g(k - k\tilde{a})a + g(k\tilde{a})(1-a)]^2} \\
&= -\frac{a(1-a)[g'(k - k\tilde{a})g(k\tilde{a})(k - k\tilde{a}) - g(k - k\tilde{a})g'(k\tilde{a})k\tilde{a}]}{-ka(1-a)[g'(k\tilde{a})g(k - k\tilde{a}) + g'(k - k\tilde{a})g(k\tilde{a})] - [g(k - k\tilde{a})a + g(k\tilde{a})(1-a)]^2}.
\end{aligned}$$

Note that the numerator converges to zero as $k \to 0$, whereas the denominator converges to $-g(0)^2 < 0$ as $k \to 0$. Hence, $\lim_{k\to 0} k\frac{d\tilde{a}}{dk} = 0$. □

II. Suppose Example 1 with different levels of k. As in the main text, if $k > 2$, there are two SEPs in which $\tilde{a}^{low} = \frac{1}{2}\left(1 - \sqrt{1 - \frac{4}{k^2}}\right)$ and $\tilde{a}^{high} = \frac{1}{2}\left(1 + \sqrt{1 - \frac{4}{k^2}}\right)$. Note that $k\tilde{a}^{low} = \frac{1}{2}\left(k - \sqrt{k^2 - 4}\right)$. By taking the derivative with respect to $k > 2$, we obtain:

$$\frac{dk\tilde{a}^{low}}{dk} = \frac{1}{2}\left(1 - \frac{k}{\sqrt{k^2-4}}\right) = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1 - \frac{4}{k^2}}}\right) < 0.$$

Because the agent's effort is strictly increasing in $k\tilde{a}^{low}$, the agent lowers his effort for a higher $k > 2$ in this case.

In the proof of Proposition 2 Part III, we showed that for any $\epsilon \in (0, 1]$, if k is sufficiently large, a low SEP $\tilde{a}^{low}$ is smaller than $\frac{\epsilon}{k}$. Since $k\tilde{a}^{low} < k \cdot \frac{\epsilon}{k} = \epsilon$ and the agent's effort is characterized by $c'(e) = k\tilde{a}^{low}$ in this case, his effort approaches 0 as $k \to \infty$. □

III. In the proof of Proposition 2 Part III, we showed that for any $\epsilon \in (0, 1]$, if k is sufficiently large, a high SEP $\tilde{a}^{high}$ is larger than $\frac{k-\epsilon}{k}$. Since $k\tilde{a}^{high} > k \cdot \frac{k-\epsilon}{k} = k - \epsilon$, the agent's effort approaches 1 as $k \to \infty$.

In what follows, we show that $k\frac{d\tilde{a}^{high}}{dk}$ is positive for any sufficiently large $k > 0$. It implies that $\lim_{k\to\infty} \frac{dk\tilde{a}^{high}}{dk} = \lim_{k\to\infty} \tilde{a}^{high} + \lim_{k\to\infty} k\frac{d\tilde{a}^{high}}{dk} > 0$, and hence, there exists a sufficiently large $\bar{k} > 0$ such that $\frac{dk\tilde{a}^{high}}{dk} > 0$ for any $k > \bar{k}$.

Let denote $f(\tilde{a}^{high}, k) = G_a(\tilde{a}^{high}) - \tilde{a}^{high} = 0$. Note that we have $\frac{\partial f}{\partial \tilde{a}^{high}} = G_a'(\tilde{a}^{high}) - 1 < 0$ by the definition of SEP. As in Part I of this proposition, by the Implicit Function Theorem, we obtain

$$k\frac{d\tilde{a}^{high}}{dk} = \frac{a(1-a)}{1 - G_a'(\tilde{a}^{high})} \cdot \frac{g'(k - k\tilde{a}^{high})g(k\tilde{a}^{high})(k - k\tilde{a}^{high}) - g(k - k\tilde{a}^{high})g'(k\tilde{a}^{high})k\tilde{a}^{high}}{[g(k - k\tilde{a}^{high})a + g(k\tilde{a}^{high})(1-a)]^2}.$$

Hence, $k\frac{d\tilde{a}^{high}}{dk} > 0$ if and only if the numerator of the above equation is positive.

By the assumption, $\frac{dg(x)x}{dx} = g'(x)x + g(x) \leq 0$ for any sufficiently large x. Since $k\tilde{a}^{high} >$

$k \cdot \frac{k-\epsilon}{k} = k - \epsilon$, $g(k\tilde{a}^{high}) \leq -g'(k\tilde{a}^{high})k\tilde{a}^{high}$ for sufficiently large k. Then:

$$\begin{aligned}
&g'(k - k\tilde{a}^{high})g(k\tilde{a}^{high})(k - k\tilde{a}^{high}) - g(k - k\tilde{a}^{high})g'(k\tilde{a}^{high})k\tilde{a}^{high} \\
\geq &g'(k - k\tilde{a}^{high})g(k\tilde{a}^{high})(k - k\tilde{a}^{high}) + g(k - k\tilde{a}^{high})g(k\tilde{a}^{high}) \\
= &g(k\tilde{a}^{high})[g'(k - k\tilde{a}^{high})(k - k\tilde{a}^{high}) + g(k - k\tilde{a}^{high})]. \qquad (2)
\end{aligned}$$

Note that $\frac{dg(x)x}{dx} = g'(x)x + g(x)$ is positive for sufficiently small $x \geq 0$. Because $k - k\tilde{a}^{high} < k - k \cdot \frac{k-\epsilon}{k} = \epsilon$, by taking a sufficiently small $\epsilon > 0$, (2) is positive for any sufficiently large k — implying $k\frac{d\tilde{a}^{high}}{dk} > 0$. □

Proof of Proposition 7.

I. If the agent is not fragile, from any seed self-view, his resulting level of SEP is the same. Hence, he chooses $e = 0$ in either case. □

II. Given $\tilde{a}_{-1} + \epsilon$, the agent chooses either $e = 0$ or $e = \tilde{a}^u - \tilde{a}_{-1} - \epsilon$. The agent exerts positive effort if and only if $\epsilon < \tilde{a}^u - \tilde{a}_{-1}$ and $\tilde{a}^{high} - \tilde{a}^{low} - c(\tilde{a}^u - \tilde{a}_{-1} - \epsilon) > 0$. Hence, we obtain $\overline{\epsilon} = \tilde{a}^u - \tilde{a}_{-1}$ and $\underline{\epsilon} = \tilde{a}^u - \tilde{a}_{-1} - c^{-1}(\tilde{a}^{high} - \tilde{a}^{low})$.

As shown in Proposition 3 Part III and Proposition 5, $\frac{d\tilde{a}^{high}}{da} > 0$, $\frac{d\tilde{a}^{low}}{da} > 0$, and $\frac{d\tilde{a}^u}{da} < 0$. Hence, for any specification of $\tilde{a}_{-1}$ stated in the proposition, $\overline{\epsilon} = \tilde{a}^u - \tilde{a}_{-1}$ is strictly decreasing in a.

Proof of Proposition 8.

By Proposition 5, fixing a_0, a high SEP is more stable when a_1 is larger. It implies $\tilde{a}_1^{high} - \tilde{a}_1^u > \tilde{a}_0^{high} - \tilde{a}_0^u$.

(i) By the assumption that the agent is not willing to exert effort in a low e-SEP, given that his initial belief is $(0, \tilde{a}_0^{low})$, he does not exert effort. Hence, $(0, \tilde{a}_0^{low})$ is an e-SEP.

(ii) By the assumption that the agent exerts effort if and only if the SEP is high and stable to a mood shock, given that the agent's initial belief is $(1, \tilde{a}_1^{high})$, he prefers to exert effort if and only if his self-esteem does not collapse after exerting effort, i.e., $\tilde{a}_1^{high} - \tilde{a}_1^u \geq \epsilon$.

(iii) By the assumption that the agent exerts effort if and only if the SEP is high and stable to a mood shock, given that the agent's initial belief is $(0, \tilde{a}_0^{high})$, he prefers not to exert effort if and only if his self-esteem would collapse after exerting effort, i.e., $\tilde{a}_0^{high} - \tilde{a}_0^u < \epsilon$. $\square$