

Wittman, Donald; Singh, Nirvikar

**Working Paper**

**Evolution and depression**

Working Paper, No. 704

**Provided in Cooperation with:**

University of California Santa Cruz, Economics Department

*Suggested Citation:* Wittman, Donald; Singh, Nirvikar (2012) : Evolution and depression, Working Paper, No. 704, University of California, Economics Department, Santa Cruz, CA

This Version is available at:

<https://hdl.handle.net/10419/98627>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

## EVOLUTION AND DEPRESSION

Donald Wittman and Nirvikar Singh

August 30, 2012

### ABSTRACT:

The standard evolutionary explanation for depression is that being emotionally depressed is adaptive. We argue that *being* depressed is not adaptive (indeed, quite the opposite), but that the *threat* of depression for bad outcomes and the *promise* of pleasure for good outcomes are adaptive because they motivate people toward undertaking effort that increases fitness. We first model the optimal emotional incentive structure. We employ a principal-agent model, where the principal is the gene and the agent is the individual. The principal-agent model is a useful construct to characterize the long run tendency of evolutionary forces to reward those characteristics that increase fitness and survival of the gene. A key difference between our setup and the standard principal-agent model is that both punishment (depression) and reward (elation) have a fitness cost to the principal. We then discuss suboptimal outcomes, including bipolar disorder, unipolar depression, and lack of motivation.

JEL: D01, D03, B52

Key words: Depression, evolution, bipolar disorder, motivation, adaptation

## EVOLUTION AND DEPRESSION

Donald Wittman and Nirvikar Singh

August 30, 2012

The widespread existence of depression across cultures is a puzzle for evolutionary theory. The prevalence of any trait, especially one that appears during reproductive years, should increase fitness. But the listlessness that often accompanies depression appears to be anything but adaptive. A number of authors have proposed a solution to this conundrum by arguing that being depressed improves fitness. Watson and Andrews (2001) and Andrews and Thomson (2002) argue that depression allows one to focus on the problem at hand, thereby enabling the depressed person to find good solutions. Hagen (2003) argues that depression signals that the person needs help, which increases reproductive advantage by shifting the burden from the depressed person to others. Stevens and Price (2000) provide still a different signaling explanation. They argue that the submissiveness in depression signals to others of higher rank that the depressed person is not a threat to them, thereby preventing the depressed person from engaging in costly challenges to dominant figures. More generally, Nesse (2000) argues that being in a depressed state is useful in that it discourages individuals from undertaking risky behavior when the payoff is likely to be negative.<sup>1</sup>

This paper argues that *being* depressed is not adaptive (indeed, quite the opposite), but that the *threat* of depression for bad outcomes and the *promise* of pleasure for good outcomes are adaptive because they motivate people toward undertaking effort that increases fitness. The threat of physical pain discourages people from putting their hands too close to a fire; the threat of emotional pain encourages individuals to undertake action that reduces the likelihood of depression, and the promise of emotional reward encourages people to undertake action that increases the likelihood of elation. We are by no means the first to suggest this parallel between physical pain and emotional pain. See for example, Thornhill and Thornhill (1989) and Nesse

---

<sup>1</sup> There are many variations on the theme that being in a depressed state is adaptive. For surveys, see Gilbert (2005) and Allen and Badcock (2006).

(2000) who use the analogy. Thornhill and Thornhill concentrate their chapter on what outcomes are most likely to bring on depression (those that are most closely associated with reduced fitness, such as a child dying) and Nesse devotes only a couple of sentences on the analogy before going on to the other explanations mentioned in the introductory remarks. But the literature does not develop a comprehensive theory of depression as a deterrent and there is a big difference between saying that less fit outcomes will result in greater depression and developing a model where the conditions for this actually being the case are shown to hold.

Depression is painful with little benefit from being in the depressed state. Here, we part company with those who believe that depression leads to increased rumination and thus increased fitness. When a person is depressed, there is increased rumination, but it is a rumination of a destructive kind, such as dwelling on suicide or on how life is without meaning. There is considerable empirical evidence that depression reduces one's ability to focus and find solutions. Austin et al. (1992) and Tsourtos et al. (2002) have shown that cognitive functioning is impaired when a person is in a depressed state, with reductions in processing speed, memory, learning and the ability to change the focus of attention (see Nettle, 2004, for an extended review of the arguments and counter-arguments). More to the point, if depression itself made things better for the person, then it would not serve as a deterrent and, counter to what we observe, people might desire to be depressed as this would make them better off in the future. And if depression could be easily dismissed, it would not be a credible threat, and therefore no longer be effective.

Part of our paper makes use of a principal-agent model, where the gene is the principal and the individual is the agent of the gene. The gene uses punishments and rewards to motivate the individual toward greater reproductive fitness. The principal-agent model is a useful construct to characterize the long run tendency of evolutionary forces to reward those characteristics that increase fitness and survival of the gene. All principal-agent models have a family resemblance. This paper differs from all of the previous work in that we consider depression. The key difference (in terms of the formal modeling) is not that there are emotional punishments (depression), but that punishments are costly both to the gene and to the individual. This is in contrast to the principal-agent literature where punishments are costless transfers to the principal. Although it does not deal with the subject of depression, Rayo and Becker's 2007 article is the

closest of those papers that are concerned with the evolution of preferences.<sup>2</sup> In their model, individuals are limited in their ability to perceive differences and the gene is limited in the intensity of rewards, but rewards do not have a fitness cost.<sup>3</sup> In our model, both punishments and rewards have a fitness cost, but we do not deal with issues of perception. Our model focuses on motivating effort, which has a positive monotonic effect on fitness. The utility cost of effort is treated as a given (our reasoning will be provided later). In the Rayo-Becker model the utility function is a blank slate and the gene creates incentives for the individual to make a certain choice,  $x^*$ , where deviations in either direction reduce fitness. We are trying to answer different questions (e.g., why people are lazy) from those raised and answered in the Rayo-Becker paper, and so it is not surprising that we have undertaken a different approach, and, given the present state of biological understanding, it is not clear which approach is closer to the biological facts. Nevertheless, we believe that a basic insight of our model, that depression and elation have fitness costs, can enhance a variety of existing models. In the appendix, we show how costly incentives can be incorporated into the Rayo-Becker model. In terms of technique, we follow closely in the footsteps of Holden (2008) who introduced the use of monotone comparative statics to the principal-agent problem. Monotone comparative statics avoids the need for concavity that plagues the first-order approach employed in the earlier principal-agent literature.

We start with an evolutionary explanation for event-based depression. Our view is that to understand chronic depression (major depressive disorder), one must first understand the reasons for event-based depression. Later in Section D, we will consider chronic depression and other failures in the incentive system. We now consider the incentive structure in greater detail so that the fundamental relationships can be readily elucidated.

---

<sup>2</sup> For a comprehensive survey of the evolutionary foundations of preferences, see Robson and Samuelson (2012). Gondolfi et al. (2002) and De Fraja (2009) argue that utility is based on sexual selection. In this paper, we consider fitness more generally and do not discuss sexual selection in particular.

<sup>3</sup> They do not consider punishments. Later in our paper, we will highlight other differences.

## A. The incentive structure

Natural selection means that those individuals who are motivated to undertake behavior that increases fitness will become more prevalent than those individuals who are not so motivated. In turn, individuals are motivated by a system of punishments and rewards. Some of these feedback mechanisms are almost completely hardwired (withdraw hand from the hot stone). Other feedback mechanisms, such as depression and elation, do not arise because pain receptors are stimulated. Instead, certain cognitive connections are made so that less desirable outcomes result in depression, while more desirable outcomes result in elation. We will label the effort undertaken to avoid depression and gain elation in time  $t$  as  $e_t$ .

Any effort requires some energy. Excess effort in and of itself causes disutility. In the absence of motivating factors, there is a natural inclination for the individual to do nothing.<sup>4</sup> While it is common in evolutionary psychology to refer to early human behavior in the savannah, the impetus to do nothing unless motivated to do otherwise is reptilian, if not earlier. At even a more basic level, engaging in physical or mental effort uses up calories, a scarce resource. Survival means that energy is conserved unless its expenditure increases fitness. We view effort cost (disutility) as the fundamental unit of account; that is,  $e_t$  is the yardstick by which emotional pleasure and pain are measured. Unless the increase in expected elation and/or the decrease in expected depression are greater than the disutility of effort, the individual will not be motivated to undertake sufficient effort.

To increase fitness, there is a need for the gene to motivate people, even at the very basic level (such as the fight or flight response).<sup>5</sup> Motivation is also required for more complicated cognitive processes that require action today for some future, possibly indirect, fitness payoff. Although we remain agnostic concerning which elements lead to reproductive success, the analysis is

---

<sup>4</sup> Because individuals are motivated, they have a hard time doing nothing.

<sup>5</sup> When we say “the gene motivates the individual toward reproductive fitness,” this is just a short-hand way of saying “those individuals who have the appropriate motivation will have more surviving offspring.”

easiest to comprehend if status is viewed as a key component of fitness in human societies.<sup>6</sup> Those who are of higher status will gain more mates if they are male and more desirable mates if they are female and both sexes will have more resources to increase the probability of their offspring surviving.<sup>7</sup> Depending on circumstances, status may depend on strength, intelligences, knowledge, bravery, etc. In turn, many of these depend on motivation of the individual. Physical skill depends not only on inherited muscular and skeletal traits, but also on training. The acquisition of knowledge depends not only on inherited intellect, but also on the time invested in learning. In turn, training and learning depend to a great degree on motivation. If a person has a great desire to be of high status (that is, the person gains great pleasure from high status and endures great pain from being low status), then the person will be motivated to undertake the costs needed to achieve higher status. And so we are back to the role of depression and elation. Finally, the culture determines what it means to be high status; there is no need for the gene to hardwire the desire to be a knight in the 14th century or a computer programmer today as long as the individual desires a higher status.

As already argued, those who are motivated to undertake effort that increases fitness will be more likely to have surviving offspring; in turn, this motivation comes from a system of punishments and rewards. Elation (reward) and depression (punishment) are based on outcomes (fitness) not inputs (effort). Parents are depressed when a child dies despite their best efforts. It is reasonable to ask why the gene does not just base punishment and reward on effort rather than on fitness. In this way, the incentives would be more closely aligned to the issue of effort. To some degree there is this incentive. It is called guilt. Guilt arises when an individual believes that a bad outcome would not have occurred if the person had acted differently. Guilt is about the individual's failure with respect to others. It appears to be later on the evolutionary ladder and is a less credible motivating device. Many individuals do not have feelings of guilt and those that

---

<sup>6</sup> Because humans and their ancestors had limited knowledge and cognitive abilities, the utility function is not simply maximization of fitness, but rather a cobbling together of various desires (have sex, satisfy a crying baby, etc.) that together approximate maximization of fitness.

<sup>7</sup> There is a very large literature on the role of status in gaining mates. For a technical example, see Cole, Mailath, and Postlewaite (1992).

do can often cognitively absolve themselves from responsibility. Guilt and pride have therefore not supplanted depression and happiness as a motivator. However, to the extent that guilt and pride exist, they, like elation and depression, are incentive devices that motivate individuals toward greater fitness.

There are two constraints on the incentive system. (1) There are chemical and biological limits on the size of the punishment and rewards. That is, the cognitive system is limited in its ability to punish and reward the individual. And (2) there are fitness costs to both punishment and rewards.<sup>8</sup> A pleasurable sensation uses up resources that might be applied elsewhere.

Furthermore, pleasure may reduce caution, which reduces fitness (a detailed discussion occurs later in the paper). The immediate effect of a person being in a depressed state is to focus on the emotional pain rather than undertaking action that improves fitness. So the punishment and rewards that incentivize fitness (and come after the fitness is determined) have fitness costs. Nevertheless, evolutionary forces mean that the net effect of punishment and/or reward on fitness is positive on average.

When individuals are at the edge of existence, which was the case for most of humanity's time on earth, it is impossible for the individual to "game the gene" and be fit. Essentially, individuals that do not undertake the maximal effort in the direction of greater fitness and instead short circuit the process and just gain pleasure without increased fitness will leave few if any offspring. And those individuals who were motivated, but in ways that decreased fitness, would also leave few offspring. Of course, in the modern world, there is considerable slack and this relationship between effort and fitness is not as tight. Nevertheless, the relationship between effort and fitness-enhancing status may still remain.

## **B. The Principal-Agent model**

We now turn to a more formal presentation. Evolution is not purposeful, but survival of the fittest can be usefully characterized in terms of a gene motivating the individual to maximize fitness.

---

<sup>8</sup> This explains why one needs both the stick of depression and the carrot of elation.



## 1. Assumptions

There are two players, an individual (agent) and a gene (the principal). But they are joined together so that the agent cannot separate from the gene and establish his/her own self, independent of the gene. The individual can only terminate the connection by terminating life.

**A1:** In time period  $t$ , there are  $n$  possible gross fitness levels,  $\theta_{1t} < \dots < \theta_{nt}$ , not including the fitness cost from either punishment or reward ( $n$  is finite). A plausible time period for  $t$  is less than one year so that factors influencing fitness fluctuate over an individual's lifetime. Note that until section B3, we consider only one period at a time.

Individual effort in time  $t$  is denoted by  $e_t$ .  $e_t \in [0, \bar{e}]$ .<sup>9</sup>

**A2:** Let the probability of  $\theta_{it}$  given  $e_t$  be  $\pi_{it}(e_t)$  with  $1 > \pi_{it}(e_t) > 0$  for all  $i$  outcomes and all  $e_t$ .

$\pi_{it}(e_t)$  is twice differentiable.  $\pi_{it}(e_t)$  has the strict *monotone likelihood ratio property*. That is,

$$\frac{\pi_{it}(e_t^H)}{\pi_{it}(e_t^L)} \text{ strictly increases as } i \text{ increases for all } e_t^H > e_t^L. \sum_{i=1}^n \pi_{it} = 1.$$

A1 and A2 can be understood as follows: In time period  $t$ , the individual is facing a set of possible (gross) fitness outcomes,  $\Theta_t = \{\theta_{1t}, \theta_{2t}, \dots, \theta_{nt}\}$ . This set of possible outcomes depends on both genetic factors such as physical strength and intelligence (but excluding the genetic basis for motivation that is considered separately) and environmental factors (such as abundant rainfall and fruit on the positive side and war and disease on the negative side) that influence fitness in time period  $t$ .  $\pi_{it}(e_t)$  depends directly on effort and ultimately on the genetic determinants of motivation ( $P_{it}$  and  $R_{it}$ ) to be introduced shortly. Increased effort increases the likelihood of high fitness outcomes and reduces the likelihood of low fitness outcomes. The overall effect of effort obeys the monotone likelihood ratio property.

---

<sup>9</sup> We treat  $e_t$  as being a scalar. We assume that all the constraints have interior points so that the Arrow-Enthoven conditions hold. To reduce clutter, we assume that the upper constraint on effort,  $\bar{e}$ , is not binding. This is reasonable because of the increasing cost to the gene from increasing punishments and rewards (see A6).

**A3:** Given  $\theta_{it}$ , the gene either rewards the agent  $R_{it} \geq 0$  or punishes the agent with  $-P_{it}$ , where  $P_{it} \geq 0$ . At times, we will use  $R_{it}$  and  $P_{it}$  instead of the longer expressions  $R_{it}(\theta_{it})$  and  $P_{it}(\theta_{it})$ . The two sets of subscripts are visual reminders that both  $\theta$  and the punishment and reward structure vary over time and outcome. Later, we will show that the individual will not be both punished and rewarded for the same fitness outcome.

**A4:** The individual's utility in time period  $t$  is  $u(e_t, R_{it}, P_{it}) = R_{it} - P_{it} - e_t$ .

Utility is in this simple form because  $R_{it}$  and  $P_{it}$  are utility and disutility, respectively, and everything is measured in effort cost.<sup>10</sup> The individual's expected utility in time period  $t$  is

$$U^t = \sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t) - e_t.$$

Let  $P_t$  and  $R_t$  be vectors of  $P_{it}$  and  $R_{it}$ , respectively.

**A5:** For each  $e_t$ , there exist non-empty sets of  $\{P_t, R_t\}$  such that  $\sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t) - e_t \geq 0$ .

The right-hand side of the inequality is the participation constraint. Here the participation constraint is the decision to continue living, which is based on expectations.<sup>11</sup> As we will see, punishment not only reduces the individual's utility, but also is costly to the gene. Therefore, unlike the standard principal-agent model, the participation constraint is unlikely to be binding. In fact, most suicides are not event-based, but due to a failure in mood regulation, where the

---

<sup>10</sup> If the reader is bothered by measuring in terms of disutility of effort, the reader could employ  $v(e_t)$  instead, where  $e_t$  is effort and  $v$  is the disutility.

<sup>11</sup> We view the decision to live as being determined by the individual's expectation about the future. In Section C, we track utility over time and show that punishment and reward adjust to circumstances so that expected future utility conforms to the left-hand side of A5, regardless of the actual outcome and resulting punishment in the present period. This does not hold for individuals who have mood disorders, such as chronic depression (see section D).

punishment-reward system is not optimally configured (see section D). Note that the individual cares about utility, rather than fitness, per se, but survival of the fittest means that individuals gain more utility from being more fit.

**A6:** There is a fitness cost,  $K^P(P_{it})$ , to punishment and a fitness cost, to reward,  $K^R(R_{it})$ :

$K^P(0) = 0; K^{P'}(P_{it}) > 0; K^{P''}(P_{it}) > 0; K^R(0) = 0; K^{R'}(R_{it}) > 0; K^{R''}(R_{it}) > 0$ .<sup>12</sup> Because these terms are positive,  $K^P(P_{it})$  and  $K^R(R_{it})$  are preceded by a minus sign when calculating *net* fitness. We assume that  $\lim_{P_{it} \rightarrow \bar{P}} K^P(P_{it}) = \infty$  and  $\lim_{R_{it} \rightarrow \bar{R}} K^R(R_{it}) = \infty$ . This means that we can ignore the constraints on the size of the punishments and rewards.

**A7:** The gene chooses  $P_{it}$  and  $R_{it}$  to maximize *expected net fitness* in time  $t$ ,

$$E[f_t] = \sum_{i=1}^n [\theta_{it} - K^P(P_{it}) - K^R(R_{it})] \pi_{it}(e_t^*), \text{ subject to}$$

$$e_t^* \in \arg \max_{e_t \in [0, \bar{e}]} \left\{ \sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t) - e_t \right\}$$

$$\sum_{i=1}^n [-P_{it}(\theta_{it}) + R_{it}(\theta_{it})] \pi_{it}(e_t^*) - e_t^* \geq 0.$$

In words, the gene chooses to maximize expected fitness net of the fitness cost of punishment and reward, subject to the individual choosing that effort level that maximizes the individual's

---

<sup>12</sup> Our modeling of punishment and reward is quite general. On the one hand, depression and elation could be two different and unrelated chemical reactions; hence the zero point. On the other hand, depression and elation could be deviations of some chemical, say dopamine, from normal levels (for most people). Then bad outcomes would result in the person being depressed and having lower than the normal level of dopamine; while good outcomes would result in the person being elated and having more than the normal level of dopamine. In either case, deviations from the norm would involve fitness costs. For those who are chronically depressed, their normal amount of dopamine would be lower than the norm.

utility, as well as being subject to the individual's participation constraint. If there is more than one level of  $e_t$  that maximizes utility, we assume that the individual chooses that level which maximizes the gene's objective function.

Implicitly, each time period  $t$  is broken down into the following sequence: (a) the set of  $\theta_{it}$  and the function  $\pi_{it}(e_t)$  are given; (b)  $P_{it}$  and  $R_{it}$ ; (c) the individual chooses an effort level,  $e_t$ ; (d) a fitness outcome takes place according to  $\theta_{it}$  and  $\pi_{it}(e_t)$ ; and (e) the individual is punished or rewarded with a resulting net fitness,  $f_t$ .<sup>13</sup> Of course, the gene does not directly determine punishment and reward each period, but rather the gene instills in the individual the sense of failure or accomplishment based on the possibilities the individual is facing during that time period.

## 2. Results

Because the individual only cares about the net amount  $D_{it} = R_{it} - P_{it}$  and because both punishment and reward are costly to the gene, it is easy to see that the gene will employ at most one of them for each outcome. That is for all  $i$  and all  $t$ ,  $R_{it} P_{it} = 0$ .

Following Grossman and Hart (1983) and Holden (2008), we proceed in two stages. First, for every possible  $e_t^*$ , we assume that the principal, in this case the gene, chooses  $R_{it}$  and  $P_{it}$  to minimize cost. In the second stage, we consider what  $e_t^*$  the gene would like to implement. Hence, we start with the following constrained maximization problem:

$$(1) \min \sum_{i=1}^n [K^P(P_{it}) + K^R(R_{it})] \pi_{it}(e_t^*),$$

subject to

---

<sup>13</sup> In our working paper, we unpack (e), the last part of the sequence. In particular, the incentive effect takes place in period  $t$ , but the cost of the punishment ( $K^P(P)$ ) or reward ( $K^R(R)$ ) takes place in period  $t+1$ . For example, because of a bad fitness outcome at the end of period  $t$ , the individual is depressed in period  $t+1$ , reducing fitness still further. This unpacking makes the analysis more complicated, but does not produce new insights.

$$\sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t^*) - e_t^* \geq \sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t) - e_t, \forall e_t \in [0, \bar{e}]$$

$$\sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t^*) - e_t^* \geq 0. \quad P_{it} \geq 0; R_{it} \geq 0.$$

The constraints are linear in  $P_i$  and  $R_i$ , the choice variables. The first constraint says that  $e_t^*$  yields at least as much utility to the individual as every other possible  $e_t$ . The second constraint is the participation constraint. The objective function is convex in  $P_{it}$  and  $R_{it}$ . Hence the Kuhn-Tucker conditions hold.

Let  $C(e_t^*)$  be the expected incentive cost to the gene for a given  $e_t^*$  (this is the first line in (1)). Sometimes, there will be no feasible solution for a particular  $e_t^*$ , in which case the cost will be assigned a very large value for that particular  $e_t^*$  so that it will not be chosen in the second stage.

In the second stage, the gene maximizes net fitness:

$$(2) \max_{e_t^*} \sum_{i=1}^n \pi_{it}(e_t^*) \theta_{it} - C(e_t^*).$$

Denote  $e_t^{**}$  as the solution to the above problem. We focus not on the solution (or set of solutions), but on the direction of change in the set of solutions as certain variables change. In so doing, we will not assume that the maximization problem is concave in  $e_t^*$ . This is because it is unlikely to be so. Indeed, we have ruled out concavity by assigning a large cost to those  $e_t^*$  that are not feasible in the first stage. So the standard implicit function approach to comparative statics does not work. Instead, we will make use of monotone comparative statics (MCS).

Before proceeding it is useful to get a feel for some concepts in MCS.

Let  $X \subset R$  and  $Y \subset R$ .

*DEFINITION:* The function  $w(x, y)$  has *increasing differences* in  $x, y$  if, for every  $x^L, x^H \in X$  such that  $x^H > x^L$  and for every  $y^L, y^H \in Y$  such that  $y^H > y^L$ , we have  $w(x^H, y^H) - w(x^L, y^H) \geq w(x^H, y^L) - w(x^L, y^L)$ . The function  $w(x, y)$  has *strictly increasing differences* in  $x, y$  if the

preceding loose inequalities are strict inequalities.

It turns out that there is a simple *test* for increasing differences. Assume that  $w(x, y)$  is twice differentiable. Then  $\frac{\partial^2 w(x, y)}{\partial x \partial y} \geq 0$  for all  $x$  and  $y$  implies that  $w$  has increasing differences in  $x$  and  $y$ , and a strict inequality implies that  $w$  has strictly increasing differences. An analogous test exists when the function is not differentiable.

Let  $S(y) = \arg \max_{x \in X} w(x, y)$ .

*Theorem 1* (Topkis): Suppose that  $w(x, y)$  has *strictly increasing differences* in  $x, y$ . Then for  $y^L, y^H \in Y$  such that  $y^L < y^H$  and for  $x^L \in S(y^L)$  and  $x^H \in S(y^H)$ ,  $x^H \geq x^L$ .<sup>14</sup>

Given complementarity, the intuition is clear. An increase in  $y$  increases the marginal productivity of  $x$ . So, clearly  $x$  will not decrease when  $y$  increases. However it is possible for  $x$  not to increase, either. This could arise when  $x$  is restricted to being an integer and the increase in marginal productivity is not sufficiently large to yield an integer increase in  $x$  (we have ruled this out because we have differentiable functions). Another possibility is that we were at a corner solution ( $x^L$  was and continues to be the maximum possible  $x$ ) and thus  $x^H = x^L$ . Therefore, we have a loose inequality rather than a strict inequality.

We now go back to our analysis. Let  $\theta_{it} = \theta_{it}(\psi)$  where an increase in  $\psi$  is a mean preserving increase in the spread of  $\theta_{it}$ . Suppose further that there are only two outcomes, 1 and 2. This means that when  $\psi$  increases,  $\theta_{2t}(\psi)$  increases and  $\theta_{1t}(\psi)$  decreases by an equivalent amount.

That is,  $\theta_{2t}'(\psi) = |\theta_{1t}'(\psi)|$ .

*Proposition 1:* Under the above assumption, an increase in  $\psi$  will lead to  $e_t^*$  weakly

---

<sup>14</sup> Note that we have subtly switched notation. Earlier  $x^H$  was a point assumed to be greater than  $x^L$ . Now we are showing that if  $x$  is an element of the set  $S(y^H)$  and labeled  $x^H$ , then  $x^H$  is indeed weakly larger than all  $x$  that are elements of the set  $S(y^L)$  and labeled  $x^L$ .

increasing.<sup>15</sup>

*Proof:*

$$\sum_{i=1}^2 \pi_{i_t}(e_t^*)\theta_{i_t}(\psi) - C(e_t^*) = \pi_{1_t}(e_t^*)\theta_{1_t}(\psi) + \pi_{2_t}(e_t^*)\theta_{2_t}(\psi) - C(e_t^*)$$

Taking the cross-partials of  $\pi_{1_t}(e_t^*)\theta_{1_t}(\psi) + \pi_{2_t}(e_t^*)\theta_{2_t}(\psi) - C(e_t^*)$  with respect to  $e_t^*$  and  $\psi$ , we get:  $\pi_{1_t}'(e_t^*)\theta_{1_t}'(\psi) + \pi_{2_t}'(e_t^*)\theta_{2_t}'(\psi) = [\pi_{2_t}'(e_t^*) - \pi_{1_t}'(e_t^*)]|\theta_{1_t}'(\psi)|$ . The first term in the brackets is positive and the second term is negative but it is multiplied by a negative so that the whole term is positive as is the absolute value term. So we have strictly increasing differences. In turn, this means an increase in  $\psi$  will lead to  $e^*$  weakly increasing by Theorem 1.///

In a nutshell, increase in the spread of outcomes will lead to more effort because there are more intense incentives *when* effort has a more positive impact on fitness. More generally, whenever  $\sum_{i=1}^n \pi_{i_t}'(e_t^*)\theta_{i_t}'(\psi) > 0$ , there will be more intense incentives. Because of the monotone likelihood ratio property, there exists a  $j$  such that  $\pi_{i_t}'(e_t^*) \geq 0$  for  $i > j$ , and  $\pi_{i_t}'(e_t^*) \leq 0$  for  $i < j$ .

Consequently, we can rewrite  $\sum_{i=1}^n \pi_{i_t}'(e_t^*)\theta_{i_t}'(\psi) > 0$  as  $\sum_{i=j+1}^n \pi_{i_t}'(e_t^*)\theta_{i_t}'(\psi) > \sum_{i=1}^j |\pi_{i_t}'(e_t^*)\theta_{i_t}'(\psi)|$ .

Let us look back at the first stage.

*Proposition 2:* For  $P_{it}, P_{jt} > 0$ , for all  $i < j$  and thus for all  $\theta_{1t} < \theta_{jt}$ , we have  $P_{it} > P_{jt}$ .

*Proof:*

$$L = \sum_{i=1}^n [K^P(P_{it}) + K^R(R_{it})]\pi_{i_t}(e_t^*) - \lambda \left[ \sum_{i=1}^n [-P_{it} + R_{it}]\pi_{i_t}(e_t^*) - e_t^* + \sum_{i=1}^n [P_{it} - R_{it}]\pi_{i_t}(e_t) + e_t \right]$$

---

<sup>15</sup> If  $e_t$  is a vector of  $m$  types of effort,  $e_{t1}, e_{t2}, \dots, e_{tm}$ , where each  $e_{ij} \in [0, \bar{e}_j] = E_j$  and  $E_j \times E_k \times \Psi$  for all  $j$  and  $k$ ,  $\frac{\partial^2 \pi_{i_t}(e_{t1}, e_{t2}, \dots, e_{tm})}{\partial e_{ij} \partial e_{ik}} > 0$  for all  $j \neq k$ , and  $\pi_{i_t}(e_t)$  has the strict *monotone likelihood ratio property* for all  $e_{ij}$ , then Proposition 1 will also hold.

$$-\mu \left[ \sum_{i=1}^n [-P_{it} + R_{it}] \pi_{it}(e_t^*) - e_t^* - 0 \right]$$

Kuhn-Tucker conditions:<sup>16</sup>

$$L_{P_{it}} = K^P{}'(P_{it}) \pi_{it}(e_t^*) + \lambda [\pi_{it}(e_t^*) - \pi_{it}(e_t)] + \mu \pi_{it}(e_t^*) \geq 0; P_{it} \geq 0; L_{P_{it}} P_{it} = 0$$

$$L_{R_{it}} = K^R{}'(R_{it}) \pi_{it}(e_t^*) - \lambda [\pi_{it}(e_t^*) - \pi_{it}(e_t)] - \mu \pi_{it}(e_t^*) \geq 0; R_{it} \geq 0; L_{R_{it}} R_{it} = 0$$

$$L_{\lambda} = \sum_{i=1}^n [P_{it} - R_{it}] [\pi_{it}(e_t^*) - \pi_{it}(e_t)] + e_t^* - e_t \leq 0; \lambda \geq 0; L_{\lambda} \lambda = 0$$

$$L_{\mu} = \sum_{i=1}^n [P_{it} - R_{it}] \pi_{it}(e_t^*) + e_t^* \geq 0; \mu \geq 0; L_{\mu} \mu = 0$$

$$\text{Hence for } P_{it} > 0, K^P{}'(P_{it}) = \lambda \left[ \frac{\pi_{it}(e_t)}{\pi_{it}(e_t^*)} - 1 \right] - \mu$$

Note that  $\left[ \frac{\pi_{it}(e_t)}{\pi_{it}(e_t^*)} - 1 \right]$  must be positive if  $P_{it}$  strictly positive. By the monotone likelihood ratio

property, for  $e_t < e_t^*$ ,  $\frac{\pi_{it}(e_t)}{\pi_{it}(e_t^*)} > 1$  and  $\frac{\pi_{it}(e_t)}{\pi_{it}(e_t^*)}$  decreases as  $i$  increases beyond  $i=1$ .  $P_{it}$  must

decrease as well, because  $K^P(P_{it})$  is a strictly convex function of  $P_{it}$ . ///

Our primary interest has been to apply the principal-agent theory to explain depression. In so doing, we have expanded somewhat the scope of the theory qua theory. In the standard principal-agent problem, punishing the agent is a monetary transfer from the agent to the principal. In contrast, depression is not a transfer from the agent to the principal as both the individual in terms of utility and the gene in terms of fitness are directly harmed by the punishment. Furthermore, because the gene and the individual are bound together, the solution of selling the right to the agent when the agent is risk neutral is not possible.

---

<sup>16</sup> The Lagrange multipliers will vary over time, but to reduce unnecessary clutter, we have suppressed the subscripts for these multipliers.



### 3. Changes in expectations overtime

So far we have only considered a one-period model. We will now expand the model to more than one period. Because we are talking about an individual, the time periods might be in terms of days, weeks, or possibly months. This is a much smaller time scale than typically used in biology, where time is in generations and fitness is measured in number of offspring or some related measure. Here we consider variation in the individual's fitness and the resulting emotional state even before the individual has children. A somewhat different concept of fitness is in order. As we will see, fitness is viewed here as being an expectation about the future which ultimately is about offspring.

The context in which the individual lives (the distribution of  $\theta_i$ ) is not fixed but varies over time due to changing cultural and environmental circumstances. And because the environment is not fixed, the incentive system cannot be a hardwired response to specific cognitive events that change more rapidly than the gene. Prehistoric man was not depressed because he did not get into Yale. We characterize this change in the distribution of  $\theta_i$  as a shift in the set of outcomes.

**A8:** Let  $\theta_{it} = \theta_{i,t-1} + M_i$ ;  $M_i > 0$  for  $i > 1$ ;  $M_1 = 0$ .

This means that the distance between the  $\theta_i$  does not change over time (but as we will see, the levels do).

**A9:** Let  $\pi_{it}(e_t) = \pi_i(e_t)$ . I.e., the values of  $\theta_i$  change over time, but not their probability functions.

A8 and A9 together mean that the marginal productivity of effort remains the same overtime.

**A10:** Let  $\theta_{1t} = \theta_{1,t-1} + f_{t-1} - f_{t-2}$

Essentially, we are assuming that an increase (decrease) in fitness in period  $t-1$  shifts the distribution of possible fitness outcomes in period  $t$  upwards (downwards). It may not be immediately apparent that the shift is as we have characterized. Therefore, below, we derive the formula. For now, all we need to know is that there is a shift in  $\theta_1$ . We recognize that, in general, change will be more complicated than the simple model we are presenting here, but we believe that it illustrates some important points and that the upward and downward shifts in the

distribution (even if not in perfect lockstep) will tend to be of first-order importance. For those who are dissatisfied with our shift assumptions (A8-A10), one could instead employ the argument that individuals do not anticipate that the incentive structure will change in the next period.<sup>17</sup>

The habituation results would be similar.

Fitness is about the long-run reproductive success of the individual/gene and therefore is akin to stock prices, which are based on expectations. The stochastic nature of  $\theta_t$  means that fitness is re-evaluated each time period within the individual's life. The following equation says that net fitness at the end of time  $t$  is equal to the *expected* net fitness at the end of time  $t+1$ .

$$\mathbf{A11:} \quad f_t = \sum_{i=1}^n [\theta_{t+1} - K^P(P_{t+1}) - K^R(R_{t+1})] \pi_{t+1}(e_{t+1}^*) = E[f_{t+1}], \text{ expected net fitness in } t+1,$$

where  $e^*$  is that effort level that results when the gene chooses punishments and rewards to maximize net fitness. That is, fitness is calculated under the assumption that the gene will continue to induce the optimal effort in the future. Figure 1 provides a simple illustration of A11.

Given our assumptions,  $\pi_{t+1}(e_{t+1}^*) = \pi_t(e^*)$ ,  $K^P(P_{t+1}) = K^P(P_t)$ , and  $K^R(R_{t+1}) = K^R(R_t)$ ,

$$\begin{aligned} \text{A11 is then equivalent to } f_t &= \sum_{i=1}^n [\theta_{t+1} - K^P(P_t) - K^R(R_t)] \pi_t(e^*) \\ &= \theta_{t+1} + \sum_{i=1}^n [M_i - K^P(P_t) - K^R(R_t)] \pi_t(e^*) = \theta_{t+1} + B = E[f_{t+1}], \end{aligned}$$

---

<sup>17</sup> This is the approach used by Rayo and Becker in dealing with incentives across time. In their model, increased effort today makes the individual less happy tomorrow. They assume that the individual ignores this decrease in happiness by arguing the following: “the average individual sharply underestimates the degree to which he will habituate to an improvement in his economic conditions.” Agent myopia makes the maximization problem of the gene much simpler because the agent *mistakenly* believes that present behavior will not alter the incentive structure in the future. We believe that our formulation (where agent behavior today does not change the incentive structure in the future) is a useful alternative because it does not require irrationality to explain habituation. See Robson and Samuelson (2010) for a discussion of individual naiveté.

where  $B$  is the value of the summation, which does not differ over time.

The net fitness at the end of period  $t$  is the *expected* net fitness at the end of period  $t+1$ . That is, fitness is based on expectations given realizations, and of course any fitness at the end of period  $t+1$  must also be consistent with expectations concerning the next period after that, etc. So  $\theta_{1,t+1}$  adjusts up and down so that the equality between  $f_t$  and  $\theta_{1,t+1} + B$  holds.

Note that  $f_{t-1} = \theta_{1,t} + B$  and that  $f_{t-2} = \theta_{1,t-1} + B$ . From this second equality, we get  $B = f_{t-2} - \theta_{1,t-1}$ . Substituting the right-hand of this equality for  $B$  in the first equality, and rearranging the equation, we get,  $\theta_{1,t} = \theta_{1,t-1} + f_{t-1} - f_{t-2}$ . This is statement A10.

Given our assumptions equations (1) and (2) can be rewritten as follows:

$$(1') \min \sum_{i=1}^n [K^P(P_{it}) + K^R(R_{it})] \pi_i(e_i^*), \text{ subject to}$$

$$\sum_{i=1}^n [-P_{it} + R_{it}] \pi_i(e_i^*) - e_i^* \geq \sum_{i=1}^n [-P_{it} + R_{it}] \pi_i(e_i) - e_i, \forall e_i \in [0, \bar{e}]$$

$$\sum_{i=1}^n [-P_{it} + R_{it}] \pi_i(e_i^*) - e_i^* \geq 0. P_{it} \geq 0; R_{it} \geq 0.$$

$$(2') \max_{e_i^*} \sum_{i=1}^n \pi_i(e_i^*) \theta_{it} - C(e_i^*).$$

$$(2'') \max_{e_i^*} \sum_{i=1}^n \pi_i(e_i^*) [\theta_{it} + f_{t-1} - f_{t-2}] - C(e_i^*) = \sum_{i=1}^n \pi_i(e_i^*) \theta_{it} + f_{t-1} - f_{t-2} - C(e_i^*)$$

We look first at equation 1'. Nothing changes when there is a shift in the set of  $\theta_i$ . Thus  $C(e_i^*)$  remains the same. Next, let us look at equation 2'. It is readily seen that the solution to 2'' is the same as the solution to 2'. That is, depression and elation adjust to expectations. To illustrate, suppose that fitness had been stable over several periods and then there were a large increase in fitness from period  $t-2$  to  $t-1$ . There would also be a large increase in happiness (utility) from  $t-2$  to  $t-1$ . But even if the individual maintained the same level of fitness in period  $t$ , the individual's happiness would decrease because the level of effort needed to maintain that level of

fitness in period  $t$  would have decreased and therefore the reward is also less. In the long run both happiness and event-based depression are fleeting as the person adjusts to new circumstances. Bad outcomes, such as an amputated leg, are initially depressing but generally less so overtime as expectations (the values of the  $\theta_i$ ) decrease. On the other side of the ledger, winning the lottery is cause for joy, but the joy decreases overtime because expectations increase.<sup>18</sup>

In the one-period analysis, the gene was maximizing the fitness of the individual. Now, the gene is facing a repeated game, but in each period, the gene will be maximizing fitness. Even though, increased fitness today increases fitness tomorrow, the gene can do no better than maximizing expected fitness today.<sup>19</sup>

Because  $\pi_i(e_t) = \pi_i(e)$ , it never pays the individual to reduce (or increase) effort in time  $t-1$  in order to increase the utility returns to effort in time period  $t$ . The marginal utility return to effort in time period  $t$  is not affected by the individual's choice of effort in time period  $t-1$ . For a given level of effort, the expected punishment and reward in time period  $t$  are the same regardless of the level of effort or outcome in time period  $t-1$ . Hence the results in the previous sections hold in the more complicated environment that we are considering here.<sup>20</sup>

#### D. SUBOPTIMAL OUTCOMES

Until now, we have focused on the optimal motivational system. But evolution does not mean that we are born perfect. Not all of us are geniuses and about 30% of children have astigmatism (Kleinstejn et al., 2003) even though better eyesight appears to improve fitness; and, with regard to the focus of this paper, there may be improper regulation of the motivation system so that

---

<sup>18</sup> See Gilbert (2006) for an extended account of this phenomenon.

<sup>19</sup> Maximizing fitness today includes the possibility of the individual storing food for tomorrow.

<sup>20</sup> Repeated games have the potential for multiple equilibria. We have modeled the genetic incentive system as being hardwired to punish and reward for performance in the present period. In this way, the multiple equilibria problem is avoided.

there is either hyper or hypo-active implementation of depression and elation. In this section, we discuss failures in the motivational system such as bipolar disorder and clinical depression.

Going back to the eye, ophthalmologists are not only able to say that certain conditions are suboptimal, but also to explain the impact of the various deviations from the optimal. If there is more curvature in the eye than optimal, then the person will be able to see close objects in focus, but not distant ones and if the curvature of the eye is less than optimal, then the reverse will be the case. In the same way, we will be able to explain how certain results differ when the incentive system is hyperactive as opposed to hypoactive.

### **1. The difficulty in achieving the right balance**

Evolution must avoid the Scylla of persistent depression, where the person is immobilized and the Charybdis of incessant euphoria, where the person tends to be reckless and subject to addiction.<sup>21</sup> And likewise, evolution must avoid the Scylla of being bipolar and the Charybdis of having no changing moods, whatsoever (these will be discussed at further length, below). One can see the difficulty in achieving the right balance by looking at the moodiness of adolescents where the internal monitoring system is a work in progress and there are great emotional swings from reckless euphoria to withdrawn depression.

Because expectations are context dependent, it is much harder to maintain the appropriate emotional equilibrium than producing an appropriately shaped eye, for example. The chemical-

---

<sup>21</sup> Although many authorities tie addiction to depression, there is considerable evidence that addiction is tied to the euphoric states. Individuals with bipolar disorder are more than twice as likely to be alcoholics as people with unipolar depression (see Sonne and Brady, 2002). Drugs and alcohol result in a short-lived euphoria for both humans and mice. The latter are not necessarily depressed. Seeing addiction in this light may lead to different therapies to reduce addiction than those that are based on the belief the drinking is caused by depression.

Opiates are a way of short-circuiting the reward system. It therefore should not be surprising that taking of opiates interferes with the incentive structure and that those who take such drugs are less likely to strive, except striving for more opiates, while they are addicted.

biological system needs to provide just the right incentive structure in a changing environment. It should not be surprising that the balance system may itself be out of balance.<sup>22</sup>

## 2. Bipolar Disorder

In this subsection, we discuss hyperactive motivation. Bipolar I disorder occurs when the person is prone to experience extended periods of extreme euphoria and at other times, extreme depression. While the initial stages of mania are sometimes romanticized, being in a manic state is not adaptive. It is generally the case that manic episodes have a short period of elevated mood followed by a more prolonged period of disorganized thoughts and behavior, often ending in suicide.

From the viewpoint of this paper, a person with bipolar has an overly powerful incentive system that undermines the person's ability to function when in the throes of the disorder, but creates a powerful incentive for extraordinary productivity, otherwise.<sup>23</sup> It is the highly productive behavior during these more or less normal periods that is adaptive, but there is the downside—the periods of mania and depression when the person is not able to function in a productive way. Note that we are not saying that bipolar disorder increases fitness (we suspect that it does not), but rather that the disorder has some partially offsetting positive benefits when not in the throes of mania or depression. Another way of seeing this is that the destructive obsession with suicide is the downside of the obsessive concentration that occurs during more normal periods. This focus may also exist in unipolar depression.<sup>24</sup>

---

<sup>22</sup> Traits influenced by many genes tend to show high levels of maladaptive genetic variance. If many genes are involved in mood regulation, then dysfunction is quite likely (see Keller and Miller, 2006).

<sup>23</sup> See Jamison (1996) for a study of creativity and bipolar disorder. There is some disagreement concerning exactly where in the mood cycle the most enduring work is done.

<sup>24</sup> This suggests the following test. Are people who are obsessed with suicidal thoughts more focused when they are in a normal state than normal people in normal states?

### 3. Laziness

To obtain a greater appreciation of the punishment and reward system, suppose that a person had no moods whatsoever, a situation that could be considered the mirror of bipolar disorder. Just as bipolar disorder is recognized as a biological illness and from the viewpoint of this paper a failure in the control of the internal incentive system, the opposite of bipolar can arise where the incentive system is underperforming and lacking a monitor. As a result, there would be insufficient motivation to act while in a "normal" state.<sup>25</sup> Some view this as a character flaw--the person is labeled lazy. Pop psychology recommendations abound, including a "tough love" approach, where the view is that until things get bad enough the person will not be motivated to change. Others view a person lacking in motivation as being depressed, but if the person is satisfied with his/her situation, the word depressed is inappropriately applied. The observer is saying that the observer would be depressed if in a similar situation, but the unmotivated person is not depressed, just not motivated. It is something akin to telling the depressed person that he/she should not be depressed because his/her objective reality is so good. That is just saying that the observer would not be depressed if facing a similar objective reality, but the observer is not the depressed person. And in a similar way, the observer is not the unmotivated person. In this view, "lazy" individuals have a weak (hypoactive) incentive system and need not be at all depressed by their situation. And once we understand this to be the case, the treatment (if there is to be a treatment in the first place) involves a totally different approach from those presently offered.

Bipolar disorder has a strong genetic component and IQ has a strong genetic component and height has a strong genetic component. Therefore, one should not be surprised that motivation is also likely to have a genetic component. Motivation is just more likely to interact with the environment--the possibilities (or lack of possibilities) that the person faces.

---

<sup>25</sup> It should be noted that marijuana is a mood stabilizer and that increased use of marijuana is associated with reduced motivation. See Syed et al. (1991).

## 4. Unipolar Depression

### A. Event-based depression

Depression arises when outcomes are below expectations. We are by no means the first to observe this relationship. However, the earlier evolutionary explanations argue that being depressed is good for you rather than seeing the threat of depression as an incentive device. We have argued that depression is the stick that encourages individuals to undertake those strategies that are most likely to lead to greater fitness. But in order for the stick to work, failure needs to result in depression. The problem with Price's explanation and the other explanations presented at the beginning of this essay is that they only focus on the bad outcomes (in particular, depression) not on the whole picture. By looking only at depression and not its incentive effects on the behavior of the person when not depressed, these explanations miss the underlying evolutionary rationale for the punishment and reward system. It is something akin to looking at people in prison and arguing either that they are being rehabilitated to become more productive members of society (prison, like depression, is good for those who are imprisoned) or that having prisons is bad because prisoners are not productive while being incarcerated. These arguments, like the arguments that only focus on people in depressed states, ignore the incentive effects of imprisonment (depression) and the possibility that some, but not all, will be deterred from committing a crime (or from underperforming) in the first place. And all these earlier Darwinian explanations for depression miss the other half of the incentive system – elation.

### B. Chronic depression (major depressive disorder)

Because we have been looking for evolutionary-based explanations, almost the entire paper has been devoted to event-based depression. Most people do not suffer from a major depressive disorder; nevertheless, a significant number of people do. Our view is that chronic depression does not increase fitness; instead, it is a failure in achieving the right biological-chemical balance in a very delicate system that should respond to certain cognitively understood situations, but not others. Individuals who suffer from chronic depression are often immobilized for long periods of time, lack libido, and have a much higher rate of suicide than those who do not suffer from chronic depression or bipolar disorder. So being chronically depressed clearly does not increase



fitness.<sup>26</sup> According to the theory presented here, when these individuals are not suffering from depression, they are likely to be very productive, but unlikely to be so productive as to outweigh the lack of productivity during the periods of depression. In a nutshell, being chronically depressed does not enhance fitness.

## **E. AMYGDALA**

To generate further insight, we now take a brief look inside the brain. The anterior cingulate cortex, the hippocampus and the amygdala have all been viewed as the sources of depressive feelings. Here, we concentrate on the role of the amygdala. In general, increased activation of the amygdala is associated with depression.<sup>27</sup> This should be seen in the light of the amygdala's central role in directing attention by influencing cortical arousal and increased sensory and perceptual processing (see Davis and Whalen, 2001). It therefore should not be surprising that activation of the amygdala is associated with anxiety-based depression (see Davidson et al., 2009).

But why are people anxious in the first place? They are anxious because they are concerned about the outcome, in particular, the potential for physical or emotional pain. If they did not care about the outcome, they would not be anxious. A certain level of anxiety is not only a sign of motivation, but also helpful in that it makes the person more alert when needed and therefore more fit. However, as we have already seen in discussing bipolar disorder, the motivating system

---

<sup>26</sup> De Catanzaro (1984) argues that suicide by people with negative marginal product increases inclusive fitness. Like the other explanations presented in the introduction, this explanation is clever, but incomplete. While it is true that those who are 70 years and older (and therefore are the most likely to have negative marginal product) have the highest rate of suicide of any age group, the rate is higher for those who are widowers than for those who are still married, where the issue of inclusive fitness would be more prevalent. Furthermore, depression without suicide reduces inclusive fitness.

<sup>27</sup> The hyperactivity of the amygdala appears to be coupled with diminished responsiveness during depressive episodes of regions involved in emotion regulation such as the dorsal anterior cingulate and the prefrontal cortex (see Stoll et al., 2000).

can be hyperactive, in which case it is debilitating, as is the case for excessive anxiety. And speaking of bipolar disorder with its excessive punishment and rewards, it should not be at all surprising that people who suffer from this condition are very likely to be anxious. In their study, Chen and Dilsaver (1995) found that among subjects with bipolar, the lifetime prevalence of panic disorder was 20.8%; among subjects with unipolar depression, it was 10.0%; and among comparison subjects, it was 0.8%.

## **F. CONCLUDING REMARKS**

Success is not just about IQ (and luck of the draw regarding the environment in which the person lives) but also about motivation. Scientists readily accept that there is a genetic component to IQ and that IQ varies across individuals. In contrast, differing levels of motivation across individuals is rarely attributed to genes. This paper seeks to change that perspective by arguing that an individual's strength of motivation depends to a great extent on the individual's inherited punishment-reward system. It is this inherited emotional structure, particularly depression and elation, but also other emotions such as anger, that motivate people to act. Of course, it is the individual capabilities within the context of the particular culture that determines whether this drive is focused on intellectual, physical or other areas of achievement.

Understanding and treatment of major depressive disorder and bipolar disorder is enhanced if we first know the evolutionary basis for event-based depression. We have provided a theory of motivation that is based on cognitive punishments (in particular, depression) and rewards (elation). The potential for depression and elation motivate the individual to undertake actions that promote greater fitness. The punishment-reward system is to a great extent genetic, but it is hard to fine tune and thus may be either hypo or hyper-active. If the punishment-reward system is hyper-active, the individual suffers from the extremes of major depressive disorder and mania; if the punishment-reward system is hypo-active, the individual may lack motivation. We believe that our approach will lead to a deeper understanding of depression and will serve as the basis for an expanded research agenda on emotions and motivation.

## APPENDIX

In this appendix, we show how costly punishments and awards can be incorporated into Rayo and Becker's basic model. Because our model in the main body of the paper is considerably different from their model, we will need to introduce several new terms. We also modify the Rayo-Becker model by allowing for punishments.

(AA1) Let  $\theta = v(x) + s$ , where  $\theta$  is fitness not including the fitness cost of punishment and reward,  $x$  is some choice variable and  $s$  is a random variable, with a symmetric single-peaked twice-differentiable density function  $g(s)$ .  $E[s] = 0$ .  $v$  is a strictly concave twice differentiable function of  $x$  with a maximum at  $x^*$ .  $v(x)$  is symmetric around  $v(x^*)$ . Note that  $x$  unlike  $e$  has no inherent disutility. Note further that unlike effort, which always has a positive effect on fitness,  $v'(x) < 0$  for  $x > x^*$ .

(AA2)  $R(\theta)$  is the punishment (if  $R < 0$ ) or reward (if  $R > 0$ ).

(AA3) There is a lower and upper limit on  $R$ :  $\underline{R} \leq R \leq \bar{R}$ ;  $\underline{R} = -\bar{R}$ .

(AA4) The agent maximizes  $\int R(v(x) + s)g(s)ds$

(AA5) If  $|E[R|x_1] - E[R|x_2]| < \delta$ , then the person cannot rank these choices and they are equally likely. The individual chooses an  $x$  from the satisficing set  $[\underline{x}, \bar{x}]$ , where  $E[R|x^*] - E[R|\underline{x}] = E[R|x^*] - E[R|\bar{x}] = \delta$ . The choices have a uniform distribution within this set. Thus within the set,  $h(x) = 1/[\bar{x} - \underline{x}]$ , and outside of the set,  $h(x) = 0$ . Note that given all the symmetry,  $\bar{x} - x^* = x^* - \underline{x}$ .

(AA6) The gene minimizes  $[\bar{x} - \underline{x}]$  subject to  $E[R|x^*] - E[R|\underline{x}] = E[R|x^*] - E[R|\bar{x}] = \delta$  and  $|R| < \bar{R}$ .

*Proposition A (Rayo and Becker):* The gene chooses a value  $\hat{\theta}$ , such that for all values of  $\theta < \hat{\theta}$ ,  $R$  will equal  $\underline{R}$  and for all values of  $\theta \geq \hat{\theta}$ ,  $R$  will equal  $\bar{R}$ .  $\hat{\theta}$  solves  $g(\hat{\theta} - v(x^*)) = g(\hat{\theta} - v(\bar{x})) = g(\hat{\theta} - v(\underline{x}))$ .

We will not reproduce their proof. However, we will provide some intuition. The gene would like the satisficing set,  $[\underline{x}, \bar{x}]$ , to be as small as possible. Essentially, (AA5) says the distance between  $\underline{x}$  and  $\bar{x}$  gets smaller as the expected utility difference increases. This is accomplished by making punishment as negative as possible ( $\underline{R}$ ) and making the reward as positive as possible ( $\bar{R}$ ).

We will now insert our approach into the Rayo-Becker model. We drop (AA3) and instead make the following assumption:

(AA3') Let the fitness cost of punishment and reward,  $K(R)$  be increasing in  $|R|$  and symmetric; that is,  $K(R) = K(-R)$ ,  $K(0) = 0$ , and  $K'(R) > 0$ . For  $R > 0$ , assume that  $K''(R) > 0$ .

Let  $\underline{x} = x^* - z(\delta)$ ;  $\bar{x} = x^* + z(\delta)$ . Then

$$E[R | x^*] - E[R | \bar{x}] = \int R(v(x^*) + s)g(s)ds - \int R(v(x^* + z) + s)g(s)ds = E[R | x^*] - E[R | \underline{x}] = \delta$$

We substitute the following for (AA6):

(AA6') The gene maximizes  $\int_{-z}^z v(x^* + w)[1/2z]dw - \int_{-z}^{+z} \left( \int R(v(x^* + w) + s)g(s)ds \right) [1/2z]dw$

$$\text{subject to } \int R(v(x^*) + s)g(s)ds - \int R(v(x^* + z) + s)g(s)ds = \delta .$$

$$\int R(v(x^*) + s)g(s)ds - \int R(v(x^* - z) + s)g(s)ds = \delta$$

This is a more complicated than the objective function in the Rayo-Becker model.

To gain insight, we start with the solution outlined in *Proposition A*: for all values of  $\theta < \hat{\theta}$ ,  $R$  will equal  $\underline{R} = -\bar{R}$  and for all values of  $\theta \geq \hat{\theta}$ ,  $R$  will equal  $\bar{R}$ . The fitness cost of punishment and reward will then be equal to  $K(\bar{R})$ . Suppose that we stick with this solution, but increase  $\bar{R}$ . This is possible since we have eliminated the constraint on the maximum possible value of  $R$ . Then the fitness cost,  $K(\bar{R})$ , will increase at an increasing rate as we have assumed that  $K''(R) > 0$ .

The question then becomes whether the assumption that  $R$  is a constant for all values greater than  $\hat{\theta}$  makes sense in the context of our version. We first show that a shift up or down will increase cost. Suppose for example that for all values of  $\theta < \hat{\theta}$ ,  $R$  now equals 0 and for all values of  $\theta \geq \hat{\theta}$ ,  $R$  now equals  $2\bar{R}$ . We can do this as there are no longer restrictions on the size of  $R$ . The differential in expected utility given  $x^*$  and given  $\bar{x}$  remains the same. However, the total cost of punishment and reward increases because  $K''(R) > 0$ .

Let us next look at the constraints in (AA6):

$$\int_{\hat{\theta}-v(x^*)}^{\hat{\theta}-v(x^*)} R(v(x^*)+s)g(s)ds + \int_{\hat{\theta}-v(x^*)} R(v(x^*)+s)g(s)ds$$

$$- \int_{\hat{\theta}-v(x^*+z)}^{\hat{\theta}-v(x^*+z)} R(v(x^*+z)+s)g(s)ds - \int_{\hat{\theta}-v(x^*+z)} R(v(x^*+z)+s)g(s)ds = \delta$$

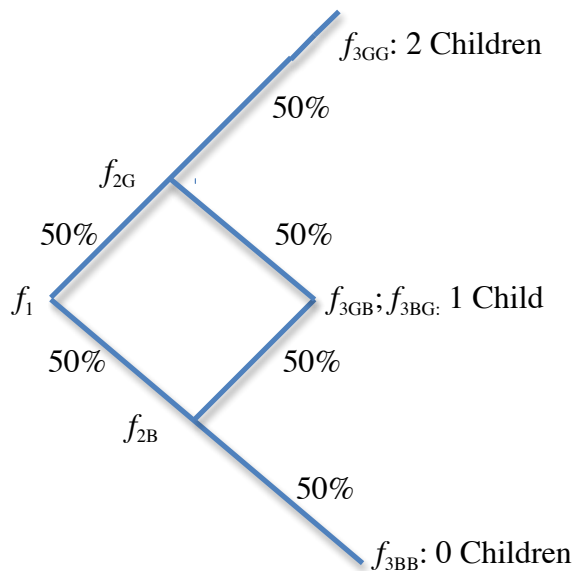
If  $R$  is fixed, then the second term equals  $\int_{\hat{\theta}-v(x^*)} R(v(x^*)+s)g(s)ds = \bar{R}[1 - G((\hat{\theta} - v(x^*)))]$

where the bar now just represents the upper value of  $R$  chosen rather than representing a constraint. The question is whether we can vary  $R$  so that the expected benefit over this range is the same, but the cost of punishment is less. The answer is no because  $K(R)$  is convex. The same holds for the other three terms. So, we now have the same answer as Rayo-Becker:

If  $\theta < \hat{\theta}$ , then  $R = \underline{R}$ ; if  $\theta > \hat{\theta}$ , then  $R = \bar{R}$ . The only difference is that in our model,  $\underline{R} = -\bar{R}$

is endogenous, where the marginal increase in  $\theta$  due to an increase in  $R$  is matched by the increased fitness cost of  $R$  so that net fitness remains the same. We have already shown that the marginal cost of  $R$  increases at an increasing rate. At the same time there is maximal benefit,  $v(x^*)$ . So there will exist a solution to the gene's maximization problem (possibly at a corner).

Figure 1: A Simple Fitness Example



For purposes of illustration, let us define fitness here as number of children and for ease of exposition, let us assume that the fitness cost of punishment is equal to the fitness cost of reward  $= K$ . In this way, if one wishes to do so, one can readily determine  $\theta_i$  given  $f_i$ . At the beginning of period 1, the individual is looking for a mate. Net fitness is  $f_1$ . At the end of period 1 and the beginning of period 2, the person either has a robust (very fertile) mate ( $f_{2G}$ ; a good outcome) or has a weak mate ( $f_{2B}$ ; a bad outcome). The probability of finding a robust mate is 50%. At the end of period 2 and the beginning of period 3, the individual with a robust mate has either produced 2 children ( $f_{3GG}$ ) or only 1 child ( $f_{3GB}$ ). The probability of having 2 children given a robust mate is 50%; the probability of have 1 child given a robust mate is 50%. The individual with a weak mate has a 50% chance of 1 child ( $f_{3BG}$ ) and a 50% chance of no children ( $f_{3BB}$ ). Upward sloping lines result in elation; downward sloping lines result in depression. Assuming number of children is net fitness and working backward down the tree, we get the following fitness values:  $f_{2G} = 1.5$ ;  $f_{2B} = .5$  and  $f_1 = .5f_{2G} + .5f_{2B} = 1$ .

## REFERENCES

Allen, Nicholas B. and Paul B.T. Badcock (2006) "Darwinian models of depression: A review of evolutionary accounts of mood and mood disorders" *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 30: 815–826.

Andrews, Paul W. and J. Anderson Thomson Jr . (2009) "Depression's Evolutionary Roots" *Scientific American* August.

Austin, M.P., Ross, M., Murray, C., Ocarroll, R.E., Ebmeier, K.P., Goodwin, G.M. (1992) "Cognitive function in major depression" *Journal of Affective Disorders* 25: 21–29.

Chen Y. W. and S. C. Dilsaver (1995) "Comorbidity of panic disorder in bipolar illness: evidence from the Epidemiologic Catchment Area Survey" *Am J Psychiatry* 152: 280-282.

Cole, H., Mailath, G., and A. Postlewaite (1992) "Social Norms, savings behavior, and growth" *Journal of Political Economy* 100: 1092-1125.

Davidson, R. J., Pizzagalli, D. and Jack Nitschke (2009) "Representation and Regulation of Emotion in Depression." In Ian Gotlib and Constance Hammen (eds.) *Handbook of Depression* New York: The Guilford Press

Davis, M. and Whalen, P. J. (2001) "The amygdala: Vigilance and emotion" *Molecular Psychiatry* 6:13-34.

de Catanzaro, D. (1984) "Suicidal ideation and the residual capacity to promote inclusive fitness: A survey." *Suicide and Life-Threatening Behavior* 14:75–87.

De Fraja, Gianni (2009) "The origin of utility: Sexual selection and conspicuous consumption" *Journal of Economic Behavior & Organization* 72: 51–69

Gilbert, Daniel (2006) *Stumbling on Happiness* New York: Knopf.

Gilbert, Paul (2005) "Evolution and depression: issues and implications" *Psychological Medicine* 36: 287–297.

Gondolfi, Arthur E., Anna Sachko Gandolfi, David P. Barash (2002) *Economics as an Evolutionary Science: from utility to fitness* New Brunswick, N.J.: Transaction

Grossman, S. J., and O. D. Hart (1983) "An Analysis of The Principal-Agent Problem" *Econometrica*, 51: 7-45.

Hagen, Edward (2003) "The bargaining model of depression" in P. Hammerstein (ed.) *Genetic and Cultural Evolution of Cooperation*. Cambridge: M.I.T. Press.

Holden, Richard T. (2008) "Comparative Statics in Principal-Agent Problems" Sloan School Working Paper.

Jamison, Kay Redfield (1996) *Touched with Fire: Manic-Depressive Illness and the Artistic Temperament* New York: Free Press.

Keller Matthew C. and Geoffrey Miller (2006) "Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best?" *Behav Brain Sci.* 29: 385-404.

Kleinstei R.N., Jones L. A., Hullett S., *et al.* (2003). "Refractive error and ethnicity in children" *Arch. Ophthalmol.* 121: 1141–7.

Nesse, Randolph (2000) "Is Depression an Adaptation?" *Arch Gen Psychiatry* 57:14-20.

Nettle, Daniel (2004) "Evolutionary origins of depression: a review and reformulation" *Journal of Affective Disorders* 81: 91–102

Price John S. (1967) "The dominance hierarchy and the evolution of mental illness" *Lancet* 2: 243-246.

Rayo, Luis and Gary Becker (2007) "Evolutionary Efficiency and Happiness" *Journal of Political Economy* 115: 900–914.



Robson, Arthur and Larry Samuelson (2010) "The evolutionary optimality of decision and experienced utility" *Simon Fraser University and Yale University Working Papers*.

Robson, Arthur and Larry Samuelson (2011) "The evolutionary foundations of preferences" In Jess Benhabib, Alberto Bisin and Matthew Jackson (eds) *Handbook of Social Economics* Elsevier.

Sonne, Susan C. and Kathleen Brady (2002) "Bipolar Disorder and Alcoholism" *Alcohol and Comorbid Mental Health Disorders* 26: 103-108.

Stevens, Anthony and John Price (2002) *Evolutionary Psychiatry* London: Routledge

Stoll, A. L., Renshaw, P.F., Yurgelun-Todd, D.A., & Cohen, B.M. (2000) "Neuroimaging in bipolar disorder: What have we learned?" *Biological Psychiatry* 48: 505-517.

Syed F. Ali, Glenn D. Newport, Andrew C. Scallet, Merle G. Paule, John R. Bailey, William Slikker Jr (1991) "Chronic Marijuana Smoke Exposure in the Rhesus Monkey IV Neurochemical Effects and Comparison to Acute and Chronic Exposure to Delta-9-Tetrahydrocannabinol (THC) in Rats." *Pharmacology, Biochemistry & Behavior* 40: 677-682.

Thase, Michael E. (2009) "Neurobiological Aspects of Depression" In Ian Gotlib and Constance Hammen (eds.) *Handbook of Depression* New York: The Guilford Press

Thornhill, Randy and Nancy W. Thornhill (1989) "The evolution of psychological pain" in R. Bell and N. Bell (eds.) *Sociobiology and the Social Sciences* Lubbock: Texas Tech University Press. 73–103.

Tsourtos, G., Thompson, J.C., Stough, C. (2002) "Evidence of an early information processing speed deficit in unipolar major depression" *Psychological Medicine* 32: 259– 265.

Watson, P.J & Andrews, P.W. 2002. "Toward a revised evolutionary adaptationist analysis of depression: The social navigation hypothesis" *Journal of Affective Disorders*. 72, 1-14.