The Evolutionary Logic of Honoring Sunk Costs

Mukesh Eswaran and Hugh M. Neary

University of British Columbia

September 8, 2013

Abstract

Though economics claims that sunk costs should not figure in current decision-making, there is ample evidence to suggest that people squander resources by honoring bygones. We argue that such wastage of resources was tolerated in our evolutionary past by Nature because it served fitness-enhancing functions. In this paper, we propose and model two such functions: the first in a non-strategic setting and the second in a strategic one. In the former, we demonstrate how the honoring of sunk costs could have arisen as a commitment device that Nature found expedient when the emotional and rational centers of the brain conflict over temptations that may sabotage long-term investments. By applying this idea to the self-concept, we argue that this model provides a rationale for cognitive dissonance, a well-established phenomenon in social psychology. The strategic reason we offer for the salience of sunk costs is that it provides the producers of goods an edge in contests over their output with potential interlopers. In either scenario, we show that Nature would have hardwired a concern for bygones.

JEL: D01, D03

Acknowledgements: *We would like to thank participants of the Economic Theory Workshop at the University of British Columbia for useful comments. We also gratefully acknowledge SSHRC for grant support.*

Corresponding Author: Mukesh Eswaran. E-mail: mukesh.eswaran@ubc.ca
Tel: 604-822-4921 Fax: 604-822-5915.

1 Introduction

Economists claim that sunk costs should not figure in current decision-making since, by definition, these expenditures cannot be retrieved and therefore should be treated as bygones. But there is a large literature, both experimental and anecdotal, in which people are observed to squander resources by honoring bygones that could not possibly make them better off (see e.g. Arkes and Ayton (1999), McAffee, Mailon and Mailon (2010), Thaler (1980)). Such excess in wasting resources is unlikely to have been tolerated by Nature in the process of evolution if it did not serve some purpose to enhance fitness in our remote past. The purpose of this paper is to suggest two such functions, whereby evolutionary processes have hard-wired a concern for sunk costs in human decision-making. The first demonstrates how a hardwired concern for sunk cost can arise in a non-strategic setting, while the second does the same in a strategic setting.

Economists have not ignored the fact that sunk costs seem salient to humans. Kanodia et al (1989) and Pendergast and Stole (1996) propose a theory where reputational effects make agents unwilling to change their decisions. McAffee, Mailon and Mailon (2010) have suggested several models—involving information content, reputation or constraints on decision-making—that provide rational reasons for taking sunk costs into account. Likewise, Baliga and Ely (2011) have proposed a model in which sunk costs may acquire salience as a 'memory kludge' because they may contain information about the value of a project previously initiated but whose value has now been forgotten.

Carmichael and MacLeod (2003) have proposed a model where, in a bilateral bargaining situation, the norm of taking account of sunk costs resolves the hold-up problem that typically results in underinvestment.

In an interesting take on sunk costs, the philosopher Nozick (1993) argued that honoring sunk costs can alleviate problems of time inconsistency and temptation. John may believe that going to the theatre is a good thing, but when the time comes he cannot rouse himself to do so because it would require relinquishing whatever else seems appealing at the time. To circumvent this, John buys tickets in advance as a precommitment device. But to do this, John must believe in the first place that his future self is going to honor sunk costs.[1] Whatever the reason for the time-inconsistent preferences—say hyperbolic discounting—the honoring of sunk costs may be used to remedy the problem of time inconsistency, but Nozick offers no reason for why sunk costs are honored in the first place.

We offer two evolutionary reasons to suggest why Nature may have hardwired a proclivity in humans to honor sunks costs, reasons that differ from those offered in the articles cited above. The first reason, non-strategic in nature, is that the built-in salience of sunk costs is an adaptation to the problem of self-control. In this, we pursue the rationale for self-control problems offered by Gifford (2002). He has persuasively argued that self-control problems arise because the suggestions of the prefrontal cortex (PFC), which is a relatively recent addition to the human brain, are in conflict with

[1]See Steele (1996) for a critique of Nozick's views on sunk costs. See also Kelly (2004) for an interesting philosophical view on sunk costs.

the impulses of the older, reptilian part of the brain. The latter is the part that assigns values to objects and constitutes the motivational system (MS) that induces action. The MS is driven by sensory cues, mainly visual, and so responds to the immediacy of these present cues whereas the PFC is capable of symbolic representation and is able to weigh future benefits at a conceptual level that is not accessible to the MS. In effect, the discount rates of the MS and PFC are different, and self-control problems arise when the dictates of the two systems are in conflict.

Self-controls problems, in Gifford's view, do not arise from being naïve about oneself and not knowing what to do, but rather from not being able to do what one knows should be done.[2] Given that Nature had evolved in sequence different parts of the brain that are at times in conflict, it is reasonable to posit that it would bring about adaptations that undermine the MS when it contradicts in a fitness-reducing manner the dictates of the PFC. We argue that the proclivity to honor sunk costs is one such adaptation. We model this idea and offer it as a non-strategic reason for the observed salience of sunk costs. Further, if we identify the considerable psychological investment in the self-concept as a sunk cost—the sense of identity may well be the ultimate sunk cost at an individual level—our model has the potential to explain pervasive phenomena such as cognitive dissonance documented in social psychology [Festinger (1957), see Aronson (1997) for a review]. Cognitive dissonance appears to be a phenomenon that arises from the need to protect the self-concept, even to the point of rejecting objective knowledge

[2]This is apparently precisely the problem that confronted St. Paul: "For the good that I would, I do not: but the evil which I would not, that I do." [Romans 7:19].

and embracing illusions.

Our strategic theory of the salience of sunk costs derives from our earlier work [Eswaran and Neary (2013)]. We argue that the honoring of sunk costs facilitates better protection of property rights and this, in turn, elicits more productive investment in the face of the possibility of distributional conflict. We demonstrate that evolutionarily stable preferences will be such that producers exhibit a decided proclivity to honor sunk costs by *undervaluing* the opportunity cost of their defensive efforts which makes them more aggressive in making those defensive efforts.[3] We also demonstrate that, in the evolutionarily stable outcome, natural selection contrives that interlopers *overestimate* the opportunity cost of their appropriative efforts, and so are less aggressive in making these attempts at appropriation. The predictions of this model are consistent with the existing evidence on territoriality in the animal world. [4]

The rest of the paper is organized as follows. In the next section, we offer a model that shows how Nature may hardwire a concern for sunk costs as a response to temptations that provoke discord between the PFC and MS areas of the brain. In Section 3 we argue that by interpreting the self-concept as a sunk cost, our theory has the scope to explain persistent findings in social psychology, such as cognitive dissonance. In Section 4 we offer a strategic reason for why Nature would hardwire a proclivity to honor sunk costs. The final sections presents our conclusions.

[3]These arguments are related to the *endowment effect* which Thaler (1980, 44) at one point defines as "the underweighting of opportunity costs".

[4]See, for example, Davied (1978), Leimar and Enquist (1994), Baugh and Forester (1994), Kemp and Wiklund (2004).

2 Sunk Costs in Absence of Strategic Interactions

We first offer a simple evolutionary model that shows how Nature may have hardwired a concern for sunk costs when there are no strategic interactions between individuals. Rather, in this model the interaction is between different, internal aspects of the individual himself; Nature finds it expedient to resolve internal conflict in a fitness-enhancing manner that manifests as attention to sunk costs.

Consider a productive opportunity that offers itself to an individual that requires an upfront fixed investment, subsequently irretrievable, of K units of labor. This might be clearing away unclaimed land and planting seeds to grow a crop, for example. We shall take K to be exogenous in this section. Producing output, however, also requires a variable input, the individual's effort denoted by l. We posit that the output, Q, is given by

$$Q = A(K)l^{\alpha}; \quad 0 < \alpha < 1, \tag{1}$$

where A denotes the exogenous total factor productivity, with $A'(K) > 0$.

The biological fitness of each individual (a measure of the size of the progeny an individual leaves behind) is assumed to be linear in consumption and effort:

$$Fitness = Consumption - Effort.$$

We deem this a reasonable functional form for fitness since an increase in consumption,

in an evolutionary setting, facilitates survival and the survival of offspring, whereas effort does the opposite by using up calories. One unit of consumption contributes 1 unit to fitness; one unit of effort reduces fitness by 1 unit. The fitness maximizing effort will be the solution to

$$\max_{l} \quad A(K)l^{\alpha} - l - K, \tag{2}$$

the solution, l^0, to which is

$$l^0 = [\alpha A(K)]^{1/(1-\alpha)}. \tag{3}$$

We denote the associated fitness level by $F^0(K)$.

To examine whether Nature would hardwire a proclivity to cater to sunk costs, we proceed as follows. First, we allow an individual's preferences to deviate from fitness. In particular, we allow the disutility of effort, which we denote by γ, to deviate from its true fitness cost of 1. The value of γ is not arbitrary; it is to be determined by Nature through natural selection so to maximize the individual's fitness. The crucial question is: Would we expect natural selection to ever fix the value of γ at a value different from 1. We propose below a realistic scenario where this is indeed the case and such that $\gamma < 1$.

Our hypothesis is that the proclivity to honor sunk costs is hardwired as a response to the problem of self-control. In this regard, we base our premise on Gifford (2002), who has argued on persuasive empirical grounds that the implementation of decisions is the result to two systems within the human brain. The evolutionarily older, reptilian brain

is responsible for emotions that provide the motivation for action by placing valuations on various options, and the more recent prefrontal cortex is responsible for reasoning. The former comprises the motivational system (MS) that responds more to visible and present cues than to abstract cues, like concepts of goods to be attained in the future; the prefrontal cortex (PFC) inhibits responses to immediate cues by using symbolic representation. Self-control problems arise when the choices dictated by the MS and PFC systems are at variance. The optimal choice dictated by the PFC on consideration of future rewards and which embody lower discount rates may be sabotaged by the presence of a visible cue (temptation) that triggers the MS system, resulting in future benefits being discounted at a higher rate than by the PFC.

To incorporate temptation into our model, we suppose that immediately after the fixed investment is undertaken, a fitness enhancing temptation can appear that generates consumption with no additional effort. It is a temptation in the sense that it is immediately present and available but to embrace it requires the individual to abandon the fixed investment. Let $g(T)$ and $G(T)$ denote the density and distribution functions for the size, T, of this temptation defined over the support $[0, \infty)$, with $g(T) > 0$ everywhere over the support. We denote the average size of the temptation by $\overline{T}$.

After temptation materializes, the individual discounts the benefit from the initial investment (which still requires a variable effort) we posit that the utility function in

this scenario, U, deviates from fitness in the following manner

$$U = \delta(c - \gamma l) - K, \tag{4}$$

where δ $(0 < \delta < 1)$ is the discount factor applied to the variable effort and delayed consumption. We assume δ is determined by evolutionary concerns *extraneous* to sunk cost considerations and arises from hardwiring in the older reptilian part of the brain as opposed to that of the prefrontal cortex. In other words, δ captures the extent of the conflict between MS and PFC over the temptation. The more visual the temptation, the greater would be the disparity and the lower the value of δ. The more symbolic the temptation, the lower will the conflict be and the higher the value of δ. Note that the disutility of effort of the fixed investment is evaluated at the true fitness cost of 1 and it is only the variable component of the effort that is evaluated at γ. This is reasonable because, when investigating the salience of sunk costs, we are interested in how *ex post* expenditures of effort are evaluated in comparison to *ex ante* ones.

If the individual does not succumb to temptation and pursues her investment, she will decide on the variable effort by solving

$$\max_{l} \quad \delta(A(K)l^{\alpha} - \gamma l) - K,$$

the solution, l^*, to which is given by

$$l^* = (\alpha A/\gamma)^{1/(1-\alpha)}.$$

Note that this optimal variable effort does not depend on the discount factor δ because both this effort and the output are discounted by the same factor; allowing for a time lag between variable effort and output serves no purpose here. By substituting the solution, we obtain the indirect utility function, $V(\gamma, K)$, as

$$V(\gamma, K) = \delta(\alpha A(K))^{\rho+1}(1/\alpha - 1)/\gamma^{\rho} - K, \tag{5}$$

where $\rho \equiv \alpha/(1-\alpha)$. We denote the associated fitness level, obtained by substituting for l^* into the fitness function, by $F(\gamma, K)$:

$$F(\gamma, K) = (\alpha A(K))^{\rho+1}(1/\alpha - 1/\gamma)/\gamma^{\rho} - K. \tag{6}$$

The discount factor δ does not appear in fitness because the endogenous variable effort has no dependence on δ. Clearly, $F(\gamma = 1, K) = F^0(K)$. Furthermore, by maximizing $F(\gamma, K)$ with respect to γ, it can readily verified that the maximum is reached at $\gamma = 1$.[5] In other words, $F_\gamma(1, K) = 0$. This is not surprising: if temptations were not catered to, Nature, by tinkering with the genes, would set the value of γ at the true cost of effort.

[5]Note also that $F(\gamma, K)$ is strictly concave in γ for $\alpha + \gamma < 2$.

The individual, however, may succumb to the temptation if it is sufficiently enticing. The utility that the individual would achieve after the realization of a specific value of the temptation is

$$\max\{T - K, V(\gamma, K)\}. \tag{7}$$

From an ex ante point of view, if $\overline{T} - K > V(\gamma, K)$, the individual would achieve a higher utility by forgoing the productive opportunity and waiting for the realization of the temptation. To keep the question relevant to sunk costs, we presume that two conditions are satisfied. First, we assume that

$$\overline{T} < F^0(K), \tag{8}$$

which implies that, on average, it does not make sense to Nature to have the individual wait around for the temptation to materialize. Second, we suppose that the distribution of temptations is such that there exists some temptation of size T' satisfying

$$T - K > V(\gamma = 1, K) \quad \text{for all } T > T', \tag{9}$$

that is, were the disutility of variable effort valued at its true fitness cost, there exist sufficiently large temptations that would induce a utility-maximizing individual to abandon his fixed investment after he has undertaken it.

Define a critical size of the temptation, $T^c(\gamma, K)$, by

$$T^c(\gamma, K) - K = V(\gamma, K), \tag{10}$$

where clearly this is the temptation that leaves the individual indifferent between succumbing to it and pursuing the original investment.

It follows from the above expressions that

$$T^c(\gamma, K) = \delta(\alpha A(K))^{\rho+1}(1/\alpha - 1)/\gamma^{\rho}. \tag{11}$$

Lemma 1 *: The following properties hold for the partial derivatives of the critical temptation, $T^c(\gamma, K)$:*

(i) $T^c_\gamma(\gamma, K) < 0$; *(ii)* $T^c_\delta(\gamma, K) > 0$; *(iii)* $T^c_K(\gamma, K) > 0$; *(iv)* $T^c_{\gamma K}(\gamma, K) = \dfrac{T^c_\gamma T^c_K}{T^c} < 0.$

Part (i) of the above lemma states that an increase in the disutility of effort associated with the variable component of the investment induces abandonment of the fixed investment at lower levels of temptation. Part (ii) says that the more symbolic the temptation, the larger it needs to be to entice the individual to abandon her investment. Part (iii) follows because higher K raises the total factor productivity of the investment and so requires a higher temptation to induce its abandonment. Finally, part (iv) says that the marginal increase in this critical temptation with higher fixed investment is lower when

the disutility of variable effort is higher.

The expected fitness, $\overline{F}(\gamma, K)$, of the individual is given by

$$\overline{F}(\gamma, K) = F(\gamma, K)G(T^c(\gamma, K)) + \int_{T^c(\gamma,K)}^{\infty} (T - K)g(T)dT. \tag{12}$$

The first term in the above expression captures the contribution to fitness from scenarios where the temptation is too small to entice the individual and the second represents scenarios where the initial investment is abandoned and the temptation is embraced. Collecting terms and using (10), the derivative of this expression with respect to γ is

$$\overline{F}_\gamma(\gamma, K) = F_\gamma(\gamma, K)G(T^c(\cdot)) + [F(\gamma, K) - V(\gamma, K)]g(T^c(\cdot))T^c_\gamma(\cdot)$$

The solution to the equation $\overline{F}_\gamma(\gamma, K) = 0$ will yield the fitness-maximizing disutility of effort $\gamma^*(K)$ that Nature would contrive through natural selection.

Proposition 1: In order to maximize expected fitness, natural selection would choose the disutility of effort to be below its true fitness cost (that is, $\gamma^*(K) < 1$).

Proof: Setting the first-order derivative equal to zero allows us to rewrite the condition as

$$\frac{F_\gamma}{F - V} = -\frac{g(T^c)}{G(T^c)}T^c_\gamma \tag{13}$$

when evaluated at γ^*. Since $-T^c_\gamma$, $g(T^c)$ and $G(T^c)$ are all positive it must be true that the terms F_γ and $F - V$ have the same sign when evaluated at γ^*. This is possible

if and only if $\gamma^* < 1$. To see this note first that $F(\gamma, K)$ is strictly concave in γ and that $F_\gamma(1, K) = 0$. Therefore $F_\gamma(\gamma, K) \leq 0$ for all $\gamma \geq 1$. Second, $F(\gamma, K) - V(\gamma, K)$ is positive at $\gamma = 1$ (because utility discounts consumption whereas fitness does not); further the magnitude of $F - V$ is increasing in γ. Therefore $F - V > 0$ for all $\gamma \geq 1$. Then F_γ and $F - V$ are opposite in sign for all $\gamma \geq 1$. It follows that the first-order condition can be satisfied only at a value $\gamma^* < 1$, where $F_\gamma(\gamma^*, K)$ and $F(\gamma^*, K) - V(\gamma^*, K)$ are each positive. ■

This result demonstrates that natural selection would fix the value of the disutility of effort to be *strictly less* than the true fitness cost of effort. This is Nature's 'precommitment' device—discovered by evolutionary trial and error—that tempers the proclivity to succumb to temptation and abandon investments that, on average, enhance fitness by more. The relative undervaluation of the cost of variable effort following the fixed investment would be perceived by the rational mind as honoring sunk costs.

Now that we have seen that Nature would contrive $\gamma^* < 1$, we can inquire how the value of this γ^* depends on the extent of the fixed cost, K. Is it the case that the greater the sunk cost the lower is the disutility of effort? Unfortunately, we cannot claim this: the behavior of γ^* as a function of K can be non-monotonic. Totally differentiating the first order condition for γ with respect to K, we obtain

$$\overline{F}_{\gamma\gamma}(\gamma, K)\frac{d\gamma^*}{dK} = -\overline{F}_{\gamma K}(\gamma, K).$$

Since $\overline{F}_{\gamma\gamma} < 0$ by the assumed concavity of $\overline{F}$ in γ, it follows that the sign of the comparative static derivative of γ^* with respect to K is fully determined by the sign of the cross-partial $\overline{F}_{\gamma K}(\gamma, K)$ of the fitness function. After manipulation (shown in the Appendix) we can express this cross-partial, evaluated at the solution γ^*, as

$$\overline{F}_{\gamma K} = (F - V)g(T^c)T^c_{\gamma K}(\epsilon + 1). \tag{14}$$

Here $\epsilon \equiv (\partial(g/G)/\partial T)(T/(g/G))$ is the elasticity of the ratio $g(T)/G(T)$ with respect to T, when T is evaluated at $T^c(\gamma^*, K)$. Since $T^c_{\gamma K}$ is negative by part (iv) of Lemma 1 and $(F - V)$ is positive at the solution, it follows that $\overline{F}_{\gamma K}$, and therefore the comparative static term $\partial\gamma^*/\partial K$, is opposite in sign to the term $(\epsilon + 1)$.

The intuition for this outcome is as follows. The maximizing choice of γ^* balances two opposing effects of γ. On the one hand fitness is increased directly, through F_γ, by an increase in γ. The overall probability weight on this inframarginal fitness change is $G(T^c)$. One the other hand, change in γ changes the critical size of temptation, T^c, switching the individual between pursuing the temptation and following up on the invested effort K. The value of this switch is $(F - V)$, which is weighted by $g(T^c)T^c_\gamma$. The optimal γ^* balances these effects, which can be expressed as in (13), with the fitness effects on the LHS and the probability effects on the RHS. An exogenous change in K will result in general in a change in both these terms, leading to an induced change in γ^*. As it happens, the multiplicative way in which $A(K)$ enters the model ensures that

the left hand side term in (13) is *independent* of K. Thus K has an effect on γ^* only through its effect on the right hand side term. The sign of this effect

$$\frac{\partial}{\partial K}[\frac{-g(T^c)}{G(T^c)}T^c_\gamma]$$

depends explicitly on the sign of $(\epsilon + 1)$.

The behavior of the elasticity ϵ depends heavily on the probability density function; it may be positive or negative in general. If the density function is increasing at T^c, or is not declining too rapidly, the elasticity may be positive, or at least greater than -1 so that $\overline{F}_{\gamma K}(\gamma^*, K) < 0$ and γ^* is decreasing in K, as expected. On the other hand, if the density function is declining rapidly the elasticity may be less than -1, ensuring that γ^* is increasing in K.

As an example, if the density function $g(T)$ is lognormal, then the elasticity is very large for low values of T, and falls monotonically below -1 as T increases. This gives rise to a negative relationship $\partial\gamma^*/\partial K$ for lower values of T and a positive relationship for higher values of T. This is illustrated in Figure 1 which shows the optimal disutility of effort as a function of the size of the sunk cost for a lognormal with mean $\mu = 1$ and standard deviation $\sigma = 0.5$.

As noted, in the Figure γ^* first declines and then increases as K increases. To obtain some intuition for this behavior suppose, initially, that $g'(T^c) \approx 0$. So when K increases and therefore the smallest temptation required to entice the individual also increases,

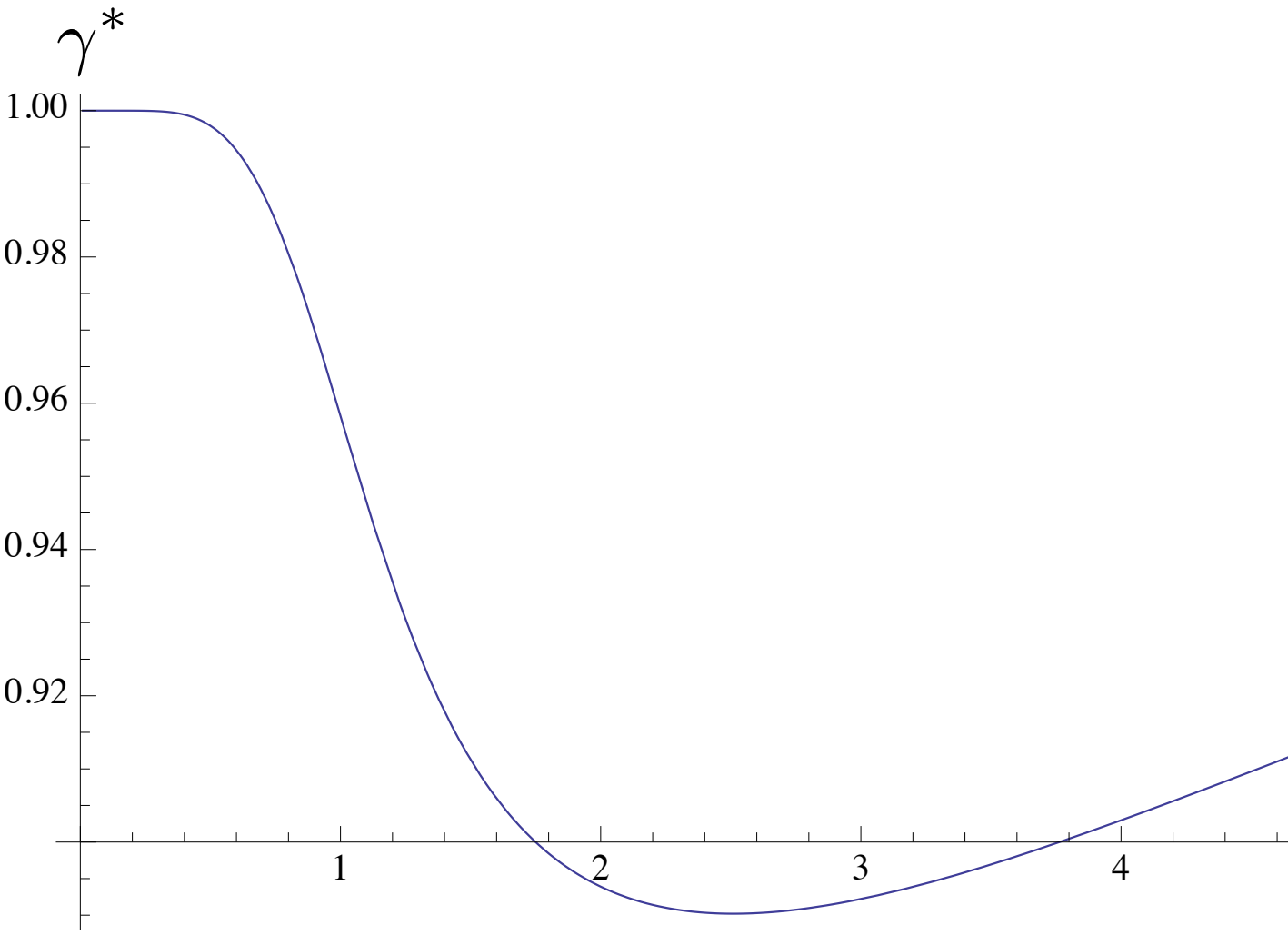

Figure 1: γ^* as a function of K ($\alpha = 0.5$; $A(K) = 3K^{1/3}$; $\delta = 0.75$)

the probability of realizing that temptation remains more or less unchanged. Since the higher K raises the total factor productivity $A(K)$, the individual gains in fitness (for some range in K) if he is dissuaded from succumbing to the temptation. To accomplish this, Nature contrives a lower γ^* to raise the size of the temptation needed to entice the individual to abandon the fixed investment. That is, the greater the sunk cost, the *lower* is the ex post disutility of effort in this case. Now suppose $g'(T^c)$ is very negative, that

is, when K increases the marginally larger T^c occurs with much lower probability.[6] In this case, Nature finds it expedient to economize on ex post variable effort by reducing the deviation of the disutility of effort from its true fitness cost since temptations higher than T^c are very low probability events. That is, Nature finds it fitness-enhancing to contrive a higher γ^*. The greater the sunk cost, the *higher* is the ex post disutility of effort in this case. This intuition suggests that Nature would have hardwired a greater proclivity to honor sunk costs (that is, hardwire a lower γ) precisely in those scenarios where diversionary temptations are potentially enticing and relatively frequent.

3 Interpretation in Terms of Self Concept

The scope and explanatory power of the simple analysis presented above is considerably enhanced if we recognize that the investment K can be interpreted more generally as one's cognitive and affective investment in the individual's self-concept. It has been recognized at least for several decades that the tendency to honor sunk costs may have something to do with the desire to maintain self-esteem [Staw (1976)]. A more recent experimental study has demonstrated that people who rated high in self-esteem measures were more desirous of being correct in their decisions and persisted in making costly

[6] This corresponds to the scenario where the elasticity ϵ of the ratio g/G is very negative; in particular, more negative than -1. Note that, since g is the density function while G is the cumulative distribution, the latter exhibits more "inertia" when T changes; consequently, the elasticity ϵ is more sensitive to marginal changes in g when T changes.

choices [Zhang and Burmeister (2006)]. It would be accurate to conceptualize self-esteem as an asset not unlike wealth—though intangible—that humans universally value greatly. Since the time of William James (1890/1981), a vast literature in psychology has documented that the pursuit of self-esteem is a deep human need and a great deal of time and money are devoted to pursuing it (see Crocker and Park (2004) for a review of this literature). This section argues that, when the self-concept is interpreted broadly as essentially a sunk cost, our framework explains one of the most pervasive phenomena of social psychology, namely, cognitive dissonance.

A person's sense of self is the individual that she perceives herself to be. It represents her notion of her identity, her deeply entrenched view of who she is and what sets her apart from others. Although there is a rudimentary sense of self in the higher primates, a well developed sense of self is only found in humans. For the most part, it is one's sense of self that is responsible for the goals one sets, the plans one makes, the investments one undertakes to achieve these goals, the precautions one takes to avoid mishaps, etc. It would not be an an exaggeration to say that the self-concept dictates most of the behavior of a person. The evolutionary benefits of the self-concept are not hard to see, since it is the over-arching governor of an individual's life. What the cognitive machinery of the individual perceives would be advantageous to survival, the self-concept utilizes to enhance the individual's survival chances [Gallup (1982)]. When exactly the sense of self evolved in humans has proved hard to pin down precisely. Some have argued that it evolved only 40,000 years ago, when modern humans (*homo sapiens sapiens*) exhibited

a great revolution in technology and art [Leary and Buttermore (2003)]. Others have argued that it may have evolved in *homo erectus*, when humans began hunting big game and the cooperation it required may have required people to self-consciously identify themselves as separate individuals [Sedikides and Skowronski (1997)].

The self-concept starts to develop at a very young age, and is observable from the age of two [Harter (2003)]. While there is an intrinsic component to the self-concept, much depends on the experiences the individual is exposed to and so there is virtually unlimited variety in the forms the self-concept might take. Though there is some malleability in the self-concept, at a given point in time it is not readily changed; it requires persistent effort to do so [Kernis and Goldman (2003)]. Couched in economic terms, the self-concept is in the nature of a stock that is the result of previous investments, conscious or unconscious. There is considerable inertia in the short run, and so the self-concept is not easily changed. This is especially the case if the required change is in the form of a perceived downgrading of self-image.

Among one of the well-documented findings of the psychology of human nature is that there is a strong bias toward maintaining a good self-image (self-esteem) of oneself. One compelling argument for this is that put forward by Leary and Downs (1995). In this argument, belonging to a group greatly facilitated survival in our evolutionary past; the consequences of being ostracized would have been very severe in harsh environments. Self-esteem, in this view, evolved as a 'sociometer'—an affective (emotional) index that responds to our perception of others' reactions to us and, therefore, is a measure of our

perception of social inclusion or exclusion. Signals of possible exclusion trigger emotional reactions (anxiety, shame, guilt, etc.) so as to motivate the experiencer to take remedial steps because those who responded to these triggers survived better.[7]

Since the self is the overseer and defender of an individual, it follows that Nature would also imbue the self with the strong sense of preservation. So not only will an individual protect herself against a physical assault—which is a biological imperative—she will also protect herself against psychological subversion because it is the self-concept that is under attack here. Since survival, not truth, is the objective of the self-concept, the self-concept will generate illusions if this enhances the chances of survival. Self requires coherence; the cognitive machinery humans are endowed with needs to make sense of the environment and to give meaning to an individual's existence and experiences [see Tesser (2003) for an overview of the literature]. As a result, normally the self cannot hold two contractory signals about the world or, more to the point here, about itself. When confronted with such a situation, individuals acquire beliefs consistent with their behavior. This is the phenomenon of 'cognitive dissonance'; inconsistent beliefs about oneself or actions that contradict beliefs produce discomfort ('dissonance') [Festinger (1957)]. Aronson (1997) has strongly argued that cognitive dissonance is related to the self-concept: people experience discomfort (dissonance) when they perform actions that are inconsistent with their self-concept.

Our theoretical result above that past effort—here reinterpreted as cognitive, affective,

[7]See Leary et al (1995) for evidence on the sociometer hypothesis.

and libidinal investment—that is pertinent to self introduces a bias towards protecting and honoring what is now a sunk cost. As a consequence, objective signals that transmit correct information (say about one's true competence or ability) may be sidelined in order to protect an existing viewpoint revolving around self because of cognitive dissonance. In other words, what appears as an irrational adherence to sunk costs stemming from the inertia implicit in the desire for a stable world view manifests as cognitive dissonance. This also provides the rationale for what has been called 'self-verification theory' in psychology, which claims that the self has a penchant to verify the correctness of its views [Swann and Read (1981)].

This still leaves unanswered the question of why Nature would want individuals to protect the self-concept through mechanisms such as cognitive dissonance when it is surely more productive to know the truth about one's self. Accurate knowledge about one's abilities and traits would prevent misguided endeavors and effort misdirected towards unattainable goals and, therefore, be fitness enhancing. What purpose is achieved by embracing illusions to protect one's bygone investment in self? One clue to this comes from the fact, recently documented, that self-esteem improves productivity; people with higher measures of self-esteem earn significantly more in the labor market after controlling for ability and human capital [Drago (2011)]. This is true even after accounting for possible reverse causality from earnings to self-esteem. The effect of self-esteem works partly by providing the individual with the confidence to enter more lucrative occupations and partly also directly (presumably) by increasing motivation. Noncognitive

skills are seen to be especially important to success in the labor market for people who are below the median in the cognitive ability scale [Lindqvist and Vestman (2011)]. Noncognitive traits such as self-esteem, therefore, are especially important for precisely the sort of people whose genes would be vulnerable to being weeded out by natural selection. The effect of self-esteem on productivity is also in keeping with the claims of Benabou and Tirole (2002), who have argued that self-confidence increases an agent's motivation when his will power is imperfect.

As [James (1890/1981)] surmised, self-esteem is not tied to one's competence in general but only to the perceived level of competence in some core areas deemed important to one's self-image. Thus there is a trade off between the loss of fitness arising from embracing falsehoods about one's core competence and the loss of productivity arising from the decline in self-esteem accompanying accurate knowledge. It is conceivable, then, that Nature would often opt for having individuals err on the side of embracing illusion rather than have them suffer loss of productivity and the attendant reduction in survival chances. This outcome, which is a resolution of the clash between the cognitive and affective aspects of self, does not arise in models based on cognition alone.[8] Perversely in this scenario, signals about one's true ability would comprise what we have called 'temptations' in the model above. It is only when the deviation from accurate knowledge is overwhelmingly large (and, therefore, dibilitating in terms of fitness) that the defenses of self against the truth will break down. Cognitive dissonance and persis-

[8]See Mischel and Morf (2003) for an overview of the networks that comprise the 'self-system' and their inherent conflicts.

tence in wrong choices such as the Concorde effect, which appear to be honoring sunk costs in the self-concept may well be a fitness-promoting route that Nature has fashioned in humans. This may be particularly so since the neural networks for emotions have been in place a lot longer—and so are more entrenched—than the cognitive machinery in humans. Therefore, the latter would have had to adapt to the former. As is well known, since mutations are typically small, natural selection is not cut out for identifying global maxima of fitness; it can only identify local maxima on the adaptive landscape [Wright (1932)]. The honoring of sunk costs in the self-concept, in other words, arises in a second-best world in which Nature is constrained by evolutionary history in its choice of the means it can invoke to further enhance fitness.

4 Sunk Costs in Strategic Interactions

The scenario we use to capture the notion of sunk cost is the following, drawn from Eswaran and Neary (2013). An individual (hereafter Player 1) applies effort and produces output. After the fact, an interloper (hereafter Player 2) may approach and attempt to steal part of the produced output. In that event there ensues a game, which we refer to as the 'distribution game', in which each player applies effort to appropriate or retain a share of the output produced by Player 1. If sunk costs are irrelevant both players, but in particular Player 1, will evaluate the cost of effort expended in the distribution game to equal the fitness cost of effort, specified as equal to 1. With

symmetry of the distribution-game effort costs both the producer and the interloper will expend equal effort and receive an equal share of the contested output. Anticipation of an equal-share outcome has implications for the incentive of the producer to expend production effort in the prior output-production stage. This production effort, output, and therefore fitness, would be increased if the producer could obtain, in the distribution game, a more-than-equal share of the output he produces. This could be achieved by Nature if an individual's marginal disutility of effort in the distribution game could evolve away from the fitness value of 1. Specifically, a reduction in the marginal disutility of distribution-game effort for the producer relative to the interloper would give the producer a better-than-equal share of the produced output, and so would give the producer an incentive to put more effort into production, raising fitness. Even though natural selection maximizes fitness, to accomplish this end Nature may contrive preferences that deviate from fitness. This sort of deviation is seen in many strategic scenarios.[9]

To model the possibilities we imagine that all individuals inherit a pair of marginal effort-disutility parameters, (γ_1, γ_2), that evolve under natural selection away from the marginal fitness cost of effort, which is 1. The disutility parameter γ_1 applies to effort expended by the individual who has the role of producer in the distribution game (Player 1), while the parameter γ_2 applies to the individual who has the role of interloper in the

[9]See, for example, Bester and Guth (1998), Bolle (2000), Ely and Yilankaya (2001), Dekel, Ely, and Yilankaya (2007), Possajennikov (2000), Schaffer (1988, 1989), Eaton and Eswaran (2003), Eswaran and Kotwal (2004), Herold and Kuzmics (2009), Eaton, Eswaran, and Oxoby (2011), and Eswaran and Neary (2013).

distribution game (Player 2). Different effort-disutility parameters will induce different expenditures of effort by producer and interloper in the distribution game, resulting in unequal sharing of the produced output. If this inequality in sharing favours the producer (as it will when $\gamma_1 < \gamma_2$) then production effort, output and therefore fitness can be increased.

In this model decision-making that involves the salience of sunk cost will appear. The producer, who has sunk an expenditure of effort into production, will be willing to expend further effort to retain that output beyond the marginal fitness cost of that effort, and beyond the marginal cost that an arbitrary third-party interloper would expend.

Given our general approach to the relevance of sunk costs, we now describe the evolutionary setting in more detail. We consider a scenario in which identical individuals live for one period, reproduce bequeathing their genes to their offspring, and then die. During this period, an individual has a probability θ of receiving a fitness-enhancing productive opportunity that requires some effort to pursue. In the prehistoric hunter-gatherer context which is the backdrop for human evolution, an example might be the appearance of a deer or other game animal which the hunter must then harvest through the expenditure of hunting effort.

The parameter θ, which is taken to be exogenous, captures the munificence of Nature. An individual who has been unlucky and has received no such productive opportunity has two options. First, he can locate (at most) one lucky individual, engage him in a distribution game and steal some of the latter's output. If, however, he does not even

manage to find a lucky individual he has to survive by opting for some low-fitness activity that offers a minimal level of consumption $\underline{c}$ that requires no effort. The probability that a producer (Player 1) will encounter an interloper is given by $\mu \equiv \min[1, (1-\theta)/\theta]$, and the probability that an unlucky individual will meet a producer is given by $\phi \equiv \min[1, \theta/(1-\theta)]$.

We assume that the output, Q, is produced by Player 1 proportional to the log of his productive effort, K :

$$Q = \alpha \log(K), \quad 0 < \alpha < 1,$$

α, the elasticity of output with respect to effort, is an inverse measure of the extent of diminishing returns to effort. Naturally, in deciding how much productive effort to apply, Player 1 will anticipate being possibly confronted by an interloper and having part of his output stolen.

When a distribution game occurs, we posit that the shares, s_1 and s_2 respectively, of the output that accrue to Players 1 and 2 are given by the symmetric functions

$$s_1 = \frac{e_1}{e_1 + e_2}; \quad s_2 = \frac{e_2}{e_1 + e_2},$$

where $e_i, i = 1, 2$, is the effort applied by Player i in the distribution game.

In the distribution game, we allow an individual's utility to deviate from fitness. Each individual is assumed to have a genetically determined pair of values for the marginal

disutility of effort, one that is relevant when the individual occupies the role of Player 1 in the distribution game, and the other that applies when the individual occupies the role of Player 2. Thus we assume that the marginal disutility of effort that governs an individual's action in the distribution game depends on the individual's role in that game as either Player 1, the producer, or Player 2, the interloper. We denote Player 1's marginal disutility of effort by γ_1 and that of Player 2 by γ_2. These are *not* chosen by individuals; they are bequeathed to them by Nature and hence are subject to selection. The crucial question we wish to settle here is whether Nature would fix these parameters (especially γ_1) at values that deviate from the marginal fitness cost of effort, which is 1.

We assume that preferences are observable; Player 1 can observe γ_2 and Player 2 can observe γ_1. In this we take our cue from much of the literature in economics cited earlier on the evolution of preferences. Darwin (1872) argued that, since not all of the facial muscles can be controlled by volition, telltale signs of untruth can be discerned. This has been confirmed by evidence in the recent psychological literature [see, for example, Porter and ten Brinke (2008)].

Thus in the distribution game, taking Q as given, Player 1 solves

$$\max_{e_1} \; u_1(e_1, e_2) := \frac{e_1}{e_1 + e_2} Q - \gamma_1 e_1,$$

and Player 2 solves

$$\max_{e_2} \; u_2(e_1, e_2) := \frac{e_2}{e_1 + e_2} Q - \gamma_2 e_2.$$

Denote the Nash equilibrium of this distribution game by $\{e_1^*(\gamma_1, \gamma_2, Q), e_1^*(\gamma_1, \gamma_2, Q)\}$. These functions are readily computed to be:

$$e_1^*(\gamma_1, \gamma_2, Q) = \frac{\gamma_2}{(\gamma_1 + \gamma_2)^2} Q \equiv \frac{s_1^* Q}{\gamma_1 + \gamma_2}; \quad e_2^*(\gamma_1, \gamma_2, Q) = \frac{\gamma_1}{(\gamma_1 + \gamma_2)^2} Q \equiv \frac{s_2^* Q}{\gamma_1 + \gamma_2},$$

where we use the equilibrium-share expressions

$$s_1^* = \frac{\gamma_2}{\gamma_1 + \gamma_2}; \qquad s_2^* = \frac{\gamma_1}{\gamma_1 + \gamma_2}.$$

The indirect utility for player 1 in the distribution game is given by

$$u_1^*(\gamma_1, \gamma_2, Q) = s_1^* Q - \gamma_1 e_1^* = (s_1^*)^2 Q.$$

For future reference note that the biological fitness of player 1 at the distribution stage is

$$f_1^*(\gamma_1, \gamma_2, Q) = s_1^* Q - e_1^* = s_1^* (1 - \frac{1}{\gamma_1 + \gamma_2}) Q.$$

The corresponding indirect utility and fitness for player 2 are

$$u_2^*(\gamma_1, \gamma_2, Q) = (s_2^*)^2 Q; \qquad f_2^*(\gamma_1, \gamma_2, Q) = s_2^* (1 - \frac{1}{\gamma_1 + \gamma_2}) Q.$$

Note first that the fitnesses of both Players 1 and 2 would be non-positive if $\gamma_1 + \gamma_2 \leq 1$. In this case, Player 2's fitness would always be improved by not entering the

distribution contest, subsisting instead on a default activity yielding $\underline{c}$. In what follows below we will restrict attention to equilibria in which $\gamma_1 + \gamma_2 > 1$.[10]

Note that s_1^* is increasing in γ_2, and decreasing in γ_1, as we would expect; and conversely, the opposite effects hold for s_2^* since $s_1^* + s_2^* \equiv 1$. Equilibrium effort for Player i is decreasing in γ_i; and it is decreasing in γ_{-i} for $\gamma_2 > \gamma_1$ (which is true in equilibrium). Finally, equilibrium efforts are both increasing in the level of output Q. The equilibrium ratio of efforts depends only on the ratio of γ's: $e_1^*/e_2^* = \gamma_2/\gamma_1$. Equilibrium shares depend on the parameters γ_i but not on Q. Finally, the equilibrium ratio of shares is given by $s_1^*/s_2^* = \gamma_2/\gamma_1$, which is greater than 1 for $\gamma_2 > \gamma_1$; this latter inequality is true in equilibrium.

We turn now to determination of the value of Q, through Player 1's choice of productive effort K. This occurs at a production stage that is prior to the distribution-game stage. Anticipating the outcome of the distribution stage, Player 1 will be challenged by an interloper with probability μ. In this case Player 1 will receive the stage 2 utility $u_1^*(\gamma_1, \gamma_2, Q)$ with probability μ. With probability $(1-\mu)$ Player 1 will not be challenged and so his post-production utility is simply Q since he dispenses no effort in the defence of his output. Thus, in choosing his productive effort, K, Player 1 will look ahead and maximize his expected utility:

$$\max_K \quad U_1 := (s_1^*)^2 \, \mu \, Q(K) + (1-\mu) \, Q(K) - K.$$

[10]This does not seem to be a burdensome restriction since the true fitness values give $\gamma_1 + \gamma_2 = 2$.

In this optimization, the player is presumed to evaluate his productive effort at its true fitness cost of 1 unit; it is only with respect to his subsequent effort in the distribution game that his marginal disutility is perceived as being γ_1, which is potentially different from 1.

The solution to this problem, $K^*(\gamma_1, \gamma_2)$, is found from the first-order condition

$$\frac{\partial U_1}{\partial K} = [(s_1^*)^2 \mu + 1 - \mu] \frac{\partial Q}{\partial K} - 1 = 0. \tag{15}$$

and is

$$K^*(\gamma_1, \gamma_2) = \alpha[(s_1^*)^2 \mu + 1 - \mu]. \tag{16}$$

As we would expect, Player 1's productive effort depends on the effort-disutilities of the two players in the distribution game, (γ_1, γ_2). These disutilities directly determine the respective efforts and equilibrium shares in the distribution game, and therefore through the value of s_1^*, impact Player 1's incentives in choosing K at the productive stage. Clearly, the optimizing value of K is increasing in player 1's ex-post share, s_1^*. In turn we have seen that s_1^* is decreasing in γ_1 and increasing in γ_2. The output in the subgame perfect equilibrium is given by $Q^*(\gamma_1, \gamma_2) = \alpha \log(K^*)$.

Finally, we can write down the expressions for the expected fitness of the two players in terms of parameters exogenous to *them*. The expected fitness, $F_1(\gamma_1, \gamma_2)$, of an

individual who finds himself in the role of Player 1 is

$$F_1(\gamma_1, \gamma_2) = \mu f_1^*(\gamma_1, \gamma_2, Q^*) + (1-\mu)Q^* - K^*$$
$$= [\mu\, s_1^* (1 - \frac{1}{\gamma_1 + \gamma_2}) + 1 - \mu]Q^* - K^*.$$

and the expected fitness, $F_2(\gamma_1, \gamma_2)$, of an individual who finds himself in the role of Player 2 is

$$F_2(\gamma_1, \gamma_2) = \phi f_2^*(\gamma_1, \gamma_2, Q^*) + (1-\phi)\underline{c} = \phi\, s_2^* (1 - \frac{1}{\gamma_1 + \gamma_2})Q^* + (1-\phi)\underline{c}.$$

Note that in the above expressions and in others, although the parameters α, θ, and $\underline{c}$ are also exogenous, the latter are suppressed for brevity and we include only γ_1 and γ_2 as arguments because these are the primary parameters of interest. The latter two parameters are also exogenous to the players, but they are endogenous in the model because they are determined by natural selection.

The manner in which we implement the working of natural selection in the determination of γ_1 and γ_2 is as follows. Nature bequeathes a pair (γ_1, γ_2) to every individual. Whether it will be γ_1 that will be relevant to the person or γ_2 will be randomly determined; if Nature offers the person a productive opportunity, then γ_1 will be relevant; if Nature does not offer such an opportunity but the person successfully becomes an interloper, then γ_2 will kick in; if the person has no opportunity even to steal, neither parameter is relevant.

Suppose all people in society have identical pairs (γ_1, γ_2) bequeathed to them. Now if a mutant appears with the pair (γ_1^m, γ_2) who achieves a higher fitness as Player 1 than all other players in the same role, then his genes will ultimately take over the population. The only scenario where γ_1 will not be replaced by a mutation is when γ_1 maximizes the fitness of Player 1, for given γ_2, that is where the value of γ_1 solves

$$\max_{\gamma_1} \qquad F_1(\gamma_1, \gamma_2).$$

Likewise, the only scenario where a mutation will not replace the genes that are relevant to people in the role of Player 2 will be when γ_2 maximizes the fitness of player 2, given the value of γ_1. That is,

$$\max_{\gamma_2} \qquad F_2(\gamma_1, \gamma_2).$$

If a pair of parameters (γ_1^*, γ_2^*) solve these two programs simultaneously, the pair would constitute a Nash equilibrium. If these values are evolutionarily stable (in a sense to be defined below), the parameter values γ_1^* and γ_2^* are deemed the values of the marginal disutilities of effort fixed by Nature for Players 1 and 2, respectively.

Consider first maximization of Player 2's fitness with respect to γ_2. The first-order condition can be written as

$$\frac{1}{\phi}\frac{\partial F_2}{\partial \gamma_2} = s_2^*(1 - \frac{1}{\gamma_1 + \gamma_2})\frac{\partial Q^*}{\partial \gamma_2} + Q^*\frac{\partial}{\partial \gamma_2}[s_2^*(1 - \frac{1}{\gamma_1 + \gamma_2})] = 0.$$

Taking the first term here to be strictly positive (i.e. $\gamma_1+\gamma_2 > 1$) ensures that the second term must be strictly negative to satisfy the (interior) first-order condition. Evaluating this latter, negative term gives

$$0 > Q^* \frac{\partial}{\partial \gamma_2}[s_2^*(1 - \frac{1}{\gamma_1 + \gamma_2})] = \frac{-Q^* \gamma_1}{(\gamma_1 + \gamma_2)^3}(\gamma_1 + \gamma_2 - 2).$$

The inequality immediately implies

$$\gamma_1 + \gamma_2 - 2 > 0.$$

Now consider maximization of player 1's fitness with respect to γ_1. The first-order condition can be written as

$$\frac{\partial F_1}{\partial \gamma_1} = \Big([\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2})+1-\mu]\frac{\partial Q}{\partial K}-1\Big)\frac{\partial K^*}{\partial s_1}\frac{\partial s_1^*}{\partial \gamma_1}+Q^*\frac{\partial}{\partial \gamma_1}[\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2})+1-\mu] = 0.$$

Using the Envelope Theorem we can substitute the expression from $\partial U_1/\partial K = 0$ (equation (15)) into the expression in parentheses in the first term on the right-hand

side to get

$$\begin{aligned}[\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2})+(1-\mu)]\frac{\partial Q}{\partial K}-[\mu(s_1^*)^2+(1-\mu)]\frac{\partial Q}{\partial K}\\ =[\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2}-s_1^*)]\frac{\partial Q}{\partial K}\\ =[\mu\, s_1^*\,(\frac{\gamma_1-1}{\gamma_1+\gamma_2})]\frac{\partial Q}{\partial K}\end{aligned}$$

In other words, using Players 1's utility maximization condition in the choice of K we can simplify the first-order fitness maximization condition as

$$\frac{\partial F_1}{\partial \gamma_1}=[\mu\, s_1^*\,(\frac{\gamma_1-1}{\gamma_1+\gamma_2})]\frac{\partial Q^*}{\partial \gamma_1}+Q^*\frac{\partial}{\partial \gamma_1}[\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2})+1-\mu]=0.$$

Since $\partial Q^*/\partial\gamma_1$ is negative it follows that the sign of the expression (γ_1-1) will be the same as the sign of the second term on the right-hand side, to satisfy the first-order condition at an interior point. Evaluating this second term gives, after simplification,

$$Q^*\frac{\partial}{\partial \gamma_1}[\mu\, s_1^*\,(1-\frac{1}{\gamma_1+\gamma_2})+1-\mu]=\frac{-\mu\, Q^*\,\gamma_2}{(\gamma_1+\gamma_2)^3}(\gamma_1+\gamma_2-2)$$

Since we know from above that $\gamma_1+\gamma_2-2$ is positive it follows that the entire term is negative in sign, and hence that (γ_1-1) is also negative in sign. Finally, combining the facts that γ_1-1 is negative and $\gamma_1+\gamma_2-2$ is positive allows us to write

$$\gamma_1<1<\gamma_2.$$

Thus we have:

Proposition 2: In any (Nash) interior equilibrium for which $\gamma_1 + \gamma_2 > 1$, the equilibrium marginal disutilities of effort satisfy $\gamma_1 < 1 < \gamma_2$.

Denote the Nash equilibrium pair by (γ_1^*, γ_2^*). Maynard Smith (1982) shows if the Nash equilibrium is unique in a finite-strategy game this equilibrium is also evolutionarily stable. We are dealing with an infinite-strategy game and so we invoke instead the concept of 'local uninvadability' to define evolutionary stability. This concept is more relevant here than Maynard Smith's.[11] No *local* producer mutant with a parameter different from γ_1^* can do better in terms of fitness than other producers in the population; likewise, no *local* interloper mutant with a parameter different from γ_2^* can do better in terms of fitness than other interlopers in the population.[12]

We cannot analytically demonstrate uniqueness of the Nash equilibrium. However, numerical solution reveals uniqueness to be the case for straightforward sets of parameter values. In these solutions the reaction function $\gamma_1^{br}(\gamma_2)$ is downward sloping: following an increase in γ_2, which makes Player 2 less aggressive as an interloper, Nature finds it expedient to decrease Player 1's disutility of effort so as to capitalize on this advantage and enable Player 1 to retain more of his output. This, in turn, provides him with greater

[11] See Definition 3 in Cressman (2009, p. 232). The Maynard Smith concept of an evolutionarily stable strategy (ESS) is defined for a finite strategy space. Generalizing the concept to a continuous strategy space is very complicated, and local uninvadability is the simplest concept in this context. See Cressman (2009) and Oechssler and Riedel (2001).

[12] Our choice of local uninvadability is consistent with Darwin's (1859) position on how natural selection operates. The conjecture that macro-mutations could bring about large changes in evolution has been dubbed a "hopeful monster", after Goldschmidt (1940). While the possibility of hopeful monsters is being raised again by a handful of researchers in the recent biology literature, our premise is firmly in the realm of what most biologists today accept as the correct view.

incentive to apply productive effort. The reaction function for $\gamma_2^{br}(\gamma_1)$, however, may slope either upwards or downwards. When resources are relatively scarce (θ less than some critical value that is greater than 0.5), $\gamma_2^{br}(\gamma_1)$ is increasing in γ_1; when resources become more abundant (θ greater than some critical value above 0.5), the best-response value for $\gamma_2^{br}(\gamma_1)$ is decreasing in γ_1.

Whatever the slope of the γ_2 reaction function, the numerical solutions all show that $\gamma_1^* < 1$ and $\gamma_2^* > 1$, consistent with Proposition 2. Player 1 thus clearly exhibits concern over his sunk cost: he discounts the disutility of effort in the post-production distribution game to well below the fitness cost of effort. Nature contrives this outcome because, by hardwiring a concern for sunk costs, it enables Player 1 to retain a bigger share of output in the distribution game, thereby giving him greater incentive to produce than he otherwise would have had. Furthermore, Nature dissuades Player 2 from too much appropriation by saddling him with a marginal disutility of effort that exceeds the marginal fitness cost of effort.

The numerical solutions also allow some comparative static relationships to be examined. In particular, the value of γ_1^* increases with the abundance of productive opportunities (as captured by θ), while the value of γ_2^* falls. As Nature becomes more munificent, the marginal disutility of Player 1 increases, while Player 2's decreases. This is because Player 1 produces more in anticipation of being challenged less frequented, and Player 2 can get away with a greater share of the output when a contest does occur. Even as θ approaches 1 the equilibrium disutility parameters do not approach the true fitness val-

ues ($\gamma_1 = 1 = \gamma_2$); rather γ_1^* converges to a value strictly less than 1 while γ_2^* converges strictly above 1. While confrontation via a post-production distribution game becomes less likely as θ increases, meaning that fewer producers are confronted by interlopers, nonetheless the possibility of conflict, and the consequent incentive impact on the level of productive effort, ensure that the disutility parameters do not converge to their true fitness value even as resource scarcity becomes insignificant.

The incentive effects generated by making sunk costs salient can be seen by considering the productive effort of player 1 in different scenarios. From (16) we know that the level of productive effort depends on the γ values and on θ. Numerical solutions demonstrate the following: (i) for any value of θ less than 1 it is true that $K^*(\gamma_1^*(\theta), \gamma_2^*(\theta), \theta) > K(1, 1, \theta)$ — productive effort is higher at the evolved disutility values than at their true fitness values; and (ii) as scarcity decreases, the levels of productive effort increase in each case, and converge to a common value as θ approaches 1. This holds even though the underlying values of the γ's do not converge as θ approaches 1. Again this underlines the point that Nature endogenously hardwires the disutilities of effort in the distributive game to make sunk costs salient to the producer, inducing thereby an increase in productive effort.

Carmichael and MacLeod (2003) present a model of *ex ante* individual investment followed by *ex post* distribution of the joint surplus from a randomly matched pair of investors. Clearly there will be an underinvestment problem *ex ante* unless the *ex post* sharing rule maintains individual investment incentives. They show that a sharing rule

based on the norm of repaying each individual's sunk costs and splitting the remaining surplus will achieve this, making sunk costs relevant in the *ex post* bargaining stage. Our model has an analogous underinvestment issue *ex ante*, which is solved, not by social norms, but by the evolution of effort-disutility parameters that make the producer more aggressive in defending his product in the *ex post* conflict.

Our outcome is also analogous to the endowment effect familiar from prospect theory. Following Thaler (1980, 43-44), treat product expropriated by the interloper as a loss to the producer, and the opportunity cost of effort expended defending that product as a foregone gain, then the endowment effect involves underweighting the importance of opportunity cost relative to direct loss. In our model consumption of the product is always valued at its fitness value, whereas in equilibrium the opportunity cost of effort is valued at $\gamma_1^* < 1$ for the producer. This relative underweighting of opportunity cost is precisely the endowment effect. Thaler (1980) takes this phenomenon as a psychological primitive; our model indicates how evolutionary forces may underlie it. In a previous paper (Eswaran and Neary (2013)) we have modelled a similar resource allocation scenario, but where the marginal utility value of consumption, rather than the marginal disutility of effort, could deviate from its fitness value. The result there is a pair of consumption valuations for player 1 and player 2, denoted (v_1, v_2), such that in equilibrium the producer values the output more than the interloper, $v_1 > v_2$. This pattern is clearly the analogue of $\gamma_1 < \gamma_2$ in the current model, and can also be interpreted directly as a form of endowment effect; we have interpreted it in psychological terms as a hardwired

version of the concept of private property.

5 Summary

We have shown plausible circumstances in which the process of evolution by natural selection might have the effect of predisposing individuals, who have already sunk effort into a project, to perceive a discounted marginal disutility for further effort needed to complete or secure the project. This hardwired undervaluation of the cost of continuation effort relative to initial effort, and, in the strategic case, relative to the perceived cost of effort to others, constitutes a Concorde or sunk cost effect. It leads to irrational behavior in the sense that sunk costs are not being ignored in decision-making; however, it is a behavior that has been rational in the evolutionary sense that it improved fitness for those exhibiting the behavior through evolutionary time.

One model considers the case of a single decision-maker who faces the possibility of a distracting temptation after effort has been sunk in a project. Conflict arises between the MS, which favors immediate gratification, and the PFC which calculates symbolically. Immediate gratification through the temptation, and associated loss of evolutionary fitness through loss of the sunk cost, will be disproportionately the outcome, unless evolution redresses the balance somewhat by making completion of the project appear less costly in effort terms than fitness alone would dictate. The result is that sunk costs are honored through excessive subsequent effort.

The second model considers the strategic situation in which a producer, having expended effort on production, may also have to expend additional effort to defend his product against an interloper. Nature may provide disutility-of-effort parameters such that the producer undervalues the cost of additional effort devoted to protecting his product, while the interloper overvalues the cost of additional effort devoted to appropriating the producer's product. In this way, the producer of a product will always appear to an outsider to be excessively willing to put effort into honouring his sunk costs of production.

In the case of both models, what might appear to the observer as a sunk cost fallacy or Concorde effect at work, is interpreted as the behavior resulting from a psychological hardwiring wrought by Nature that leads the individual to undervalue the cost of additional effort expenditure directed towards projects that involve already sunk effort.

APPENDIX

Derivation of of Equation (14):

The cross-partial derivative $\overline{F}_{\gamma K}$ is given by

$$\overline{F}_{\gamma K} = F_{\gamma K}G(T^c) + F_{\gamma}g(T^c)T^c_K + (F_K - V_K)g(T^c)T^c_{\gamma} + (F - V)(g'(T^c)T^c_{\gamma}T^c_K + g(T^c)T^c_{\gamma K}) \tag{17}$$

Note first that, by calculation,

$$F_{\gamma K} = (\rho + 1)\frac{A'(K)}{A(K)}F_{\gamma} \qquad \text{and} \qquad F_K - V_K = (\rho + 1)\frac{A'(K)}{A(K)}(F - V).$$

Substitute these values into the first and third terms of (17) and add to get

$$F_{\gamma K}G(T^c) + (F_K - V_K)g(T^c)T^c_{\gamma} = (\rho + 1)\frac{A'(K)}{A(K)}[F_{\gamma}G(T^c) + (F - V)g(T^c)T^c_{\gamma}] = 0$$

where the last equality follows because the term in brackets is the first-order condition. Thus $\overline{F}_{\gamma K}$, evaluated at γ^*, reduces to the second and fourth terms in (17). Collecting

these two terms and substituting for F_γ from the first-order condition gives

$$\begin{aligned}
\overline{F}_{\gamma K} =& F_\gamma g(T^c)T_K^c + (F-V)(g'(T^c)T_\gamma^c T_K^c + g(T^c)T_{\gamma K}^c) \\
=& (F-V)[g'(T^c)T_\gamma^c T_K^c + g(T^c)T_{\gamma K}^c - g(T^c)T_K^c] \\
=& (F-V)T_{\gamma K}^c[(g' - \frac{g^2}{G})\frac{T_K^c T_\gamma^c}{T_{\gamma K}^c} + g] \\
=& (F-V)g(T^c)T_{\gamma K}^c[\epsilon_{(g/G)} + 1]
\end{aligned}$$

where we have used the fact that $T^c = T_K^c T_\gamma^c / T_{\gamma K}^c$ in deriving the elasticity term.

References

[1] Arkes, H.R. and P. Ayton (1999), "The Sunk Cost And Concorde Effects: Are Humans Less Rational Than Lower Animals?", *Psychological Bulletin*, 125(5), pp. 591-600.

[2] Aronson, E. (1997), "Back to the Future: Retrospective Review of Leon Festinger's A Theory of Cognitive Dissonance", *American Journal of Psychology*, 110, pp. 127-157.

[3] Baugh, J.R. and D.C. Forester (1994), "Prior Residence Effect in the Dart-Poison Frog, Dendrobates Pumilio", *Behaviour*, 131, pp. 207-224.

[4] Benabou, R. and J. Tirole (2002), "Self-Confidence and Personal Motivation", *Quarterly Journal of Economics*, 117, pp. 871-915.

[5] Bester, H. and W. Guth (1998), "Is Altruism Evolutionarily Stable?", *Journal of Economic Behavior and Organization*, 34, pp. 193-209.

[6] Bolle, F. (2000), "Is Altruism Evolutionarily Stable? And Envy and Malevolence?", *Journal of Economic Behavior and Organization*, 42, pp. 131-133.

[7] Cressman, R. (2009), "Continuously Stable Strategies, Neighborhood Superiority and Two-player Games with Continuous Strategy Space", *International Journal of Game Theory*, 38, pp. 221-247.

[8] Crocker, J. and L.E. Park (2004), "The Costly Pursuit of Self-Esteem", *Psychological Bulletin*, 130, pp. 392-414.

[9] Darwin, C. (1859), *On the Origin of Species*, John Murray, London.

[10] Darwin, C. (1872), *The Expression of the Emotions in Man and Animals*, University of Chicago Press, Chicago.

[11] Davies, N.B. (1978), "Territorial Defence in the Speckled Wood Butterfly (Parage aegeria): The Resident Always Wins", *Animal Behavior*, 26, pp. 138-147.

[12] Dekel, E., J.C. Ely, and O. Yilankaya (2007), "Evolution of Preferences", *Review of Economic Studies*, 74, pp. 685-704.

[13] Drago, F. (2011), "Self-Esteem and Earnings", *Journal of Economic Psychology*, 32, pp. 480–488.

[14] Eaton, B.C. and M. Eswaran (2003), "The Evolution of Preferences and Competition: A Rationalization of Veblen's Theory of Invidious Comparisons", *Canadian Journal of Economics*, 36, pp. 832-859.

[15] Eaton, B.C., M. Eswaran, and R. Oxoby (2011), " 'Us' and 'Them': The Origin of Identity and its Economic Implications", *Canadian Journal of Economics*, 44, pp. 719-748.

[16] Ely, J. and O. Yilankaya (2001), "Nash Equilibrium and the Evolution of Preferences", *Journal of Economic Theory*, 97, pp. 255-272.

[17] Eswaran, M. and A. Kotwal (2004), "A Theory of Gender Differences in Parental Altruism", *Canadian Journal of Economics*, 37, pp. 918-950.

[18] Eswaran, M. and H.M. Neary. (2013), "An Economic Theory of the Evolutionary Origin of Property Rights.", forthcoming in *American Economic Journal: Microeconomics*.

[19] Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Evanston, IL: Row, Peterson.

[20] Gallup, G.G. (1982), "Self-awareness and the Emergence of Mind in Primates", *American Journal of Primatology*, 2, pp. 237-248.

[21] Gifford, A. (2002), "Emotion and Self-Control", *Journal of Economic Behavior & Organization*, 49, pp. 113–130.

[22] Goldschmidt, R. (1940), *The Material Basis of Evolution*, Yale University Press, New Haven.

[23] Harter, S. (2003), "The Development of Self-Representations during Childhood and Adolescence", Ch. 30 in *Handbook of Self and Identity*, (eds) M.R. Leary and J.P. Tangney, Guilford Press, New York.

[24] Herold, F., and C. Kuzmics (2009), "Evolutionary Stability of Discrimination Under Observability", *Games and Economic Behavior*, 67, pp. 542-551.

[25] James, W. (1890/1981), *Principles of Psychology*, Harvard University Press, Mass.

[26] Kemp, D.J. and C. Wiklund (2004), "Residency Effects in Animal Contests",*Proceedings: Biological Sciences*, 271, pp. 1707-1711.

[27] Kernis, M.H. and B.M. Goldman (2003), "Stability and Variability in Self-Concept", Ch. 6 in *Handbook of Self and Identity*, (eds) M.R. Leary and J.P. Tangney, Guilford Press, New York.

[28] Leary, M.R. and N.R. Buttermore (2003), "The Evolution of the Human Self: Tracing the Natural History of Self-Awareness", *Journal for the Theory of Social Behavior*, 33, pp. 365-404

[29] Leary, M.R., E.S. Tambor, S.K. Terdal, and D.L. Downs (1995), "Self-Esteem as an Interpersonal Monitor: The Sociometer Hypothesis" , *Journal of Personality and Social Psychology*, 68, pp. 518-530.

[30] Leimar, O. and M. Enquist (1984), "The Effect of Asymmetries in Owner-Intruder Interactions", *Journal of Theoretical Biology*, 111, pp. 475-491.

[31] Lindqvist, E. and R. Vestman (2011), "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment", *American Economic Journal: Applied Economics*, 3, pp. 101–128.

[32] Maynard Smith, J. (1982), *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.

[33] Mischel, W. and C.C. Morf (2003), "The Self as a Psycho-social Dynamic Processing System: A Meta-perspective on a Century of the Self in Psychology", Ch. 2 in *Handbook of Self and Identity*, (eds) M.R. Leary and J.P. Tangney, Guilford Press, New York.

[34] Oechssler, J. and F. Riedel (2001), "Evolutionary dynamics on infinite strategy spaces", *Economic Theory*, 17, pp. 141-162.

[35] Porter, S., and L. ten Brinke (2008), "Reading between the Lies: Identifying Concealed and Falsified Emotions in Universal Facial Expressions", *Psychological Science*, 19, pp. 508-514.

[36] Possajennikov, A. (2000), "On the Evolutionary Stability of Altruistic and Spiteful Preferences", *Journal of Economic Behavior and Organization*, 42, pp. 125-129.

[37] Schaffer, M.E. (1988), "Evolutionarily Stable Strategies for a Finite Population and a Variable Contest Size", *Journal of Theoretical Biology*, 132, pp. 469-478.

[38] Schaffer, M.E. (1989), "Are Profit-Maximizers the Best Survivors? A Darwinian Model of Economic Natural Selection", *Journal of Economic Behavior and Organization*, 12, 29-45.

[39] Sedikides, C. and J.J. Skowronski (1997), "The Symbolic Self in Evolutionary Context", *Personality and Social Psychology Review*, 1, pp. 80-102.

[40] Staw, B. M. (1976) "Knee-deep in the Big Muddy: A Study of Escalating Commitment to Chosen Course of Action", *Organizational Behavior and Human Performance*, 16, pp. 27-44.

[41] Swann, W. B., Jr., and S.J. Read (1981), "Acquiring Self-knowledge: The Search for Feedback that Fits", *Journal of Personality and Social Psychology*, 41, pp. 1119-1128.

[42] Tesser, A. (2003), "Self-Evaluation", Ch. 14 in *Handbook of Self and Identity*, (eds) M.R. Leary and J.P. Tangney, Guilford Press, New York.

[43] Thaler, R. (1980), "Toward a Positive Theory of Consumer Choice", *Journal of Economic Behavior & Organization*, 1(1), pp. 39-60.

[44] Wright, S. (1932), "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution", *Proceedings of the VI International Congress of Genetics*,1, pp. 356-366.

[45] Zhang, L. and R.F. Burmeister (2006), "Your Money or Your Self-Esteem: Threatened Egotism Promotes Costly Entrapment in Losing Endeavors", *Personality and Social Psychology Bulletin*, 32, pp. 881-893.