DEPARTMENT OF ECONOMICS

ISSN 1441-5429

DISCUSSION PAPER 06/06

WEAKNESS OF WILL

Elias L. Khalil[1]

ABSTRACT

The dominant view regards weakness of will an anomaly facing the standard theory of rationality. The paper argues the opposite: What is anomalous is that weakness of will is not pervasive enough. In a simple model, the paper shows that weakness of will is the dominant strategy in a game between current self and future self. This leads to the motivating question of the paper: Why is weakness of will is not pervasive—given that precommitment and punishment are not sufficiently pervasive to remedy the weakness of will? The paper argues that the answer lies in what Adam Smith calls the "propriety" mechanism: Humans demand self-respect and, hence, exercise self-command over appetites and emotions.

Key words: property of others (justice); property of future self (prudence); decision-action gap (weakness of will); mechanisms (precommitment and propriety); trust; appetites and emotions; libertarian paternalism.

JEL classification: D0

[1] Email: Elias.khalil@buseco.monash.edu.au Khalil is at the Department of Economics, Monash University, Clayton, Australia.

WEAKNESS OF WILL

WHAT IS THE QUESTION?

In futuristic Britain portrayed in the 1971 classic film "A Clockwork Orange," directed by Stanley Kubrick, Alex has a strong appetite for ultra-violence consisting of busting the heads of innocent young and old people.[2] But Alex has a stronger appetite to be out of prison. To constrain his ultra-violent appetite, Alex volunteers, for a shorter prison term, to participate in an experimental brainwashing therapy. The experiment was successful—in fact very successful. Once he was conditioned, every time the thought of violating the property of others crossed his mind, Alex underwent through uncontrollable revulsion, sickness in the stomach, and submissiveness. Alex cannot even hurt a housefly.

Similar to Alex, Ulysses and the Sirens in Homer's *Odyssey* illustrates weakness of will [Elster, 1984, p. 36]. Ulysses (Greek name, Odysseus), the king of Ithaca, was sailing back to his country after the Trojan War and after many years of frustrated wandering. As the ship approached the island where the Sirens (half woman and half bird) live, he asked to be tied with ropes to the mast of the ship and to be released only after passing the island. In such a constrained choice, he cannot (since he alone can hear because his shipmates' ears are plugged with wax) be lured by the seductive songs of the Sirens and steer the ship into the deadly rocky coast [Homer, 1977, pp. 200-201]. Ulysses in effect deprived himself of free movement while enjoying the Sirens' song (not to mention that he denied such a pleasure to the shipmates).

Alex and Ulysses adopt the same mechanism to combat their weakness of will: They impose on themselves self-punishment. In the case of Ulysses, the self-punishment involves the reduction of the budget constraint. In the case of Ales, the self-punishment involves the lowering of utility from violence. Jon Elster [1984, 2000] calls self-punishment "precommitment". Thomas Schelling [1960] calls self-punishment the offering of "hostages" which locks one in a credible behaviour in the future.

But the idea of self-punishment does not address the basic question: Why there is weakness of will to start with? Weakness of will suggests that there is a gap between decision (i.e., the optimum) and the action. What is the origin of the decision-action gap?

---

[2] The film is based on a novel of same title by Anthony Burgess.

George Ainslie [2001] explains weakness of will by invoking the idea of special discounting, viz., hyperbolic discounting. That is, agents have more intense preference of today vs. tomorrow rewards than they do of tomorrow vs. the day-after-tomorrow rewards. However, the idea of hyperbolic discounting is ultimately not an explanation but rather a description of weakness of will. Gary Becker [1996] advances a different explanation. He models addiction as a habit. As such, agents have weakness of will because, as a result of previous consumption, they have developed the taste for the prohibited good. Again, this begs the question: Why did the agent start consuming the prohibited good in the first place, especially knowing the path-dependency of the preferences?

For a satisfactory answer, the paper offers a future-looking, rational choice model that locates the source of weakness of will. Weakness of will is actually the dominant strategy between current self and future self. If so, the proposed model predicts that weakness of will should be pervasive. Thus, the anomaly which faces the standard theory of rationality is not the evidence of weakness of will, as supposed in the literature. Rather, the anomaly is the lack of the pervasiveness of weakness of will.

This leads to the main question which motivates the paper: Why is weakness of will not as widespread as predicted by the rational choice model?

To elaborate, critics of standard theory have repeatedly pointed out that the weakness of will is the Achilles' heel of the standard theory of rationality [e.g., Elster, 1984, 1999]. In light of the simple model presented here, the non-widespread of weakness of will is the actual Achilles' heel of the standard theory of rationality.

The proposed model is based on the common idea that weakness of will amounts to hurting the interest of future self, i.e., imprudence. It is also based on the common idea that imprudence is analytically similar to hurting the interest of others, i.e., injustice. Many other thinkers have drawn the parallel between imprudence and injustice as the outcome of prisoners' dilemma game. What is new about the proposed analysis is to establish that injustice is best understood as a special case of imprudence—and not *vice versa*.

The proposed entry point of imprudence as the general case, while injustice or free-riding as the special case, should help us identify the source of weakness of will. It should shift our attention from punishment or self-punishment (i.e., precommitment)—which is usually associated with injustice and free-riding—as the entry point of thinking about weakness of will. To focus on

punishment or precommitment is misleading because it cannot allow us to see that weakness of will is not caused by the lack of punishment.

The next section, section 1, discusses further why the focus on the precommitment mechanism is misleading. Section 2 sets up the question in terms of violation of the property of future self, i.e., imprudence. Section 3 sets up the question in terms of violation of the property of the other, i.e., injustice. Section 4 set up a simple model of weakness of will that encompasses imprudence and injustice. Section 5 shows the shortcoming of three major explanations of why agents refrain from violating rights and behave in trustworthy manner. In light of these shortcomings, section 6 advances the propriety mechanism as the alternative mechanism to precommitment in abridging the decision-action gap. Section 7 suggests some implications.

1. WHY IS PRECOMMITMENT MISLEADING?

Alex and Ulysses chose what Jon Elster [2000] calls "precommitment". Precommitment is a particular kind of mechanisms which Elster defines as "constraints." Agents adopt precommitment to ensure that their action corresponds to the optimum decision (in short, "decision"). Precommitment is a mechanism that sufficiently increases the costs or sufficiently decreases the utility so that it is absolutely impossible for the agent to succumb to weakness of will.

Of course, the precommitments chosen by Ulysses and Alex are Draconian—which are needed given the colossal consequences of succumbing to weakness of will. But in less extreme cases, such as abstaining from over-eating, the agent may resort to less Draconian measures such as locking the kitchen after dinner. In all these cases, precommitment amounts to increasing the cost, or lowering the benefit, of each decision involving weakness of will so that consumer surplus would be non-positive.

What about *ex ante* commitments such as the purchase of membership in a gym, when it is cheaper to pay per use when one decides three uses per week is the optimum? Such *ex ante* commitments are not precommitments because the expenditures are sunk cost. When the agent acts and visits the gym, historical expenditures should not matter (beyond the free entry afforded by the membership). So, as defined here, precommitment is the concurrent cost of action, while commitment is the sunk cost which should not matter more than the sharp declaration of what is the optimum decision.

In this sense, precommitment vis-à-vis the property rights of future self parallels the establishment of "police society" vis-à-vis the property rights of others.[3] Precommitment is similar to placing a police guard in every aisle in the supermarket or searching the clothes of *all* customers upon leaving the store. Precommitment is similar to telling someone a secret in order to establish "credible trust"—which is actually no longer trust in the sense used here but rather to establish "assured compliance" with the contract. That is, the offered secret acts as a guarantor: The teller of the secret assures the recipient that it is not in the interest of the teller to renege on the contract in the future [Schelling, 1960]. That is, the secret here acts as a hostage handed by one party so that the other party can be assured that it would not be deceived. These examples—ranging from Elster's precommitment to Schelling's hostages—rob the agent of any *ex post* choice.

In fact, the idea of "police society" as the mechanism to solve weakness of will amounts to throwing the baby with the bathwater. If offering hostages or inflicting precommitment is pervasive, trust in human affairs is trivial. But from casual empiricism, trust still figure highly in human affairs and agents do not adopt precommitments that totally extinguish temptation.

For instance, let us use the stylized fact that shoplifting (or other kinds of violation of property) is not a widespread problem in many societies. Further, let us use he stylized fact that agents everyday in such societies find the reward of shoplifting exceeds the cost (including the likelihood of getting caught). So, why is stopping such agents from shoplifting? Why is trust so pervasive? Precommitment cannot be the explanation. Obviously, people do not lock themselves so that they do not go to stores. Also, they generally, if not definitely, do not ask the store manager to accost them while shopping. So, how can we explain trust?

As one theory reviewed below suggests, agents wear "conscience" that sends painful signals if they do what they do not want to do. But this begs the question: What is the origin of conscience? It must be somehow related to welfare calculation. But how exactly it is related? So, to invoke conscience is simply *deus ex machina* explanation that begs the question.

Another theory, not reviewed here, is that agents may close the gap by resorting to either self-rationalization or self-deception. In self-rationalization, agents may lie to themselves and pretend that they had no control over the action. The prime example of self-rationalization is when Adam explained to God that Eve is to blame for violating the agreement. In self-deception, agents lie to themselves and pretend that the decision is not optimum anyway. The apotheosis of self-

[3] The idea of "police society" differs from the familiar term "police state" where the police is used to protect the current government from its critics.

deception is the fable of the fox and the sour grapes—when the fox pretended that the grapes are sour when he failed to reach them after a couple of attempts.

But it cannot be the case that *all* agents resort *all* the time to self-rationalization and self-deception. If these are the norms, there would be no problem to start with. Thus, the terms "self-rationalization" and "self-deception" would not make any sense. To illustrate, if people are liars all the time, they cannot be lying and, hence, there is no meaning to the verb "to lie." To resort to self-rationalization and self-deception, it must be the case that agents most of the time do not use them.

So, how do agents make their action coincide with their decision—where the term "decision" is used to denote the optimum decision? The paper argues that there is a hidden mechanism called "propriety" which secures that agents behave according to the optimum decision. The decision can be about the optimum allocation of time between work and leisure or between self-interest and altruism. The decision can be about the optimum property right. To keep the discussion manageable, the decision studied here concerns only intertemporal allocation with regard to property.

If so, there is a difference between propriety with property—despite the fact that they have, as Leonidas Montes [2004, ch. 4] shows skilfully, a common etymological root. As Figure 1 shows, propriety is a *mechanism* whose function is similar precommitment: Either one abridges the gap between decision, which can be the property commitment, and action.

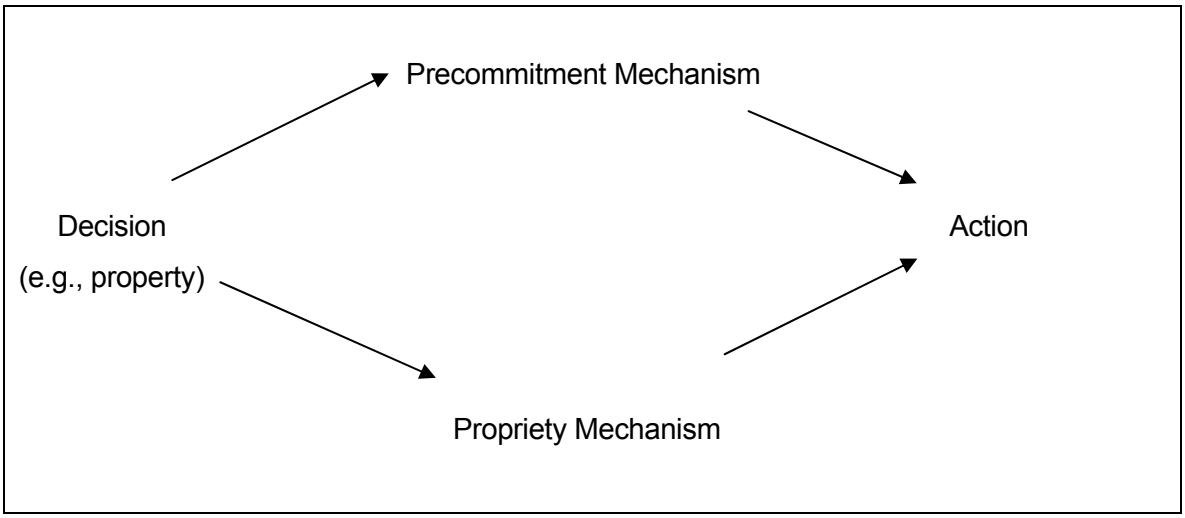

Figure 1: The action-decision Gap (weakness-of-will problem)

The decision-action gap is not recognized in economic theory based on the axioms of revealed preference. In fact, the term "choice" in economics came to obfuscate the decision-action gap—as if any choice is actually carried out. But we need to recognize the decision-action gap if we want to define weakness of will, not to mention ever identify the source of weakness of will.

2. PRUDENCE—PROPERTY OF FUTURE SELF

So defined here, prudence is the institution to which the agent is committed in order to protect his future self. To illustrate, let us examine Figure 2. The current self, and every self into the infinite tomorrows, gains 10 units of benefit if the self, e.g., smokes. The self gains good health (GH) if the self does not smoke every day in the future. And the self receives bad health (BH) if the self smokes every day in the future. These assumptions are, without changing the conclusion, relaxed in the simple model offered below.

		Future Self (infinite tomorrows)	
		Smoke	Do Not Smoke
Current Self	Smoke	10;10,…; BH	10; 0,…; GH
	Do Not Smoke	0; 10, …; BH	0; 0, …; GH

Figure 2: The Prudence Paradox

Let us assume that future self smokes in the infinite tomorrows. As a result the agent would receive bad health. What should the current self do? It is better off by 10 units to smoke today as well.

Let us assume that future self does not smoke at all in the infinite tomorrows. As a result the agent would receive good health. What should the current self do? It is better off by 10 units to smoke today.

So, current self is better off smoking today—irrespective of what future selves do. Here, the dominant strategy for current self is to smoke.

If the agent applies this reasoning everyday, the agent ends up smoking everyday, and hence, ends up with bad health. This is a suboptimal outcome, assuming that the agent enjoys more good health over the pleasure of smoking over his lifetime.

So, weakness of will is rational. Given that current smoking has no or little effect on health, it is better to smoke. But the outcome is suboptimal.

So what should the agent do to nullify the dominant strategy—given the assumption that health is superior to the pleasure of smoking? The possible mechanism discussed so far is self-punishment, i.e., what Elster calls "precommitment" or what Schelling calls "hostage." But this is not the only or even common mechanism, as discussed below.

3. JUSTICE—PROPERTY OF OTHER

As defined here, justice is the institution of fairness towards the property rights of others. But why should one respect the property rights of partners? Social contract theory, dating at least to Thomas Hobbes, has reasoned that property rights assure the best solution that avoids the suboptimal prisoners' dilemma outcome.[4] The prisoners' dilemma game has extensively been applied to public goods to illustrate how free riding leads to under-funding of public goods [Olson, 1965] and how loafing leads to the tragedy of the commons [Hardin, 1968].

I like to apply the same idea of justice as fairness to cooperation between two partners. Partners enter into a cooperative contract in order to reap the benefits of cooperation, which can be the increasing returns from division of labor or the sharing of the cost of a fixed asset such as an office, a law firm, or an apartment. One should not take advantage of one's roommate, business partner, friend, classmate, family member, marriage partner, and so on. If one free-rides, it would result in the break-up of partnership and, hence, in the loss of the benefit of cooperation.

Let us focus on partnership where there is no referee. Many cooperative venues do have referees or arbitrators. But to understand the function of the referee, we need first to analyze cooperative arrangements without referees.

Let us assume two friends agree to share an apartment and divide the chores of cleaning between them. As shown in Figure 3, the current self, and every self into the future gains 10 units of benefit if the self free rides, i.e., exerts low effort of cleaning. The agent receives bad

---

[4] The Red Queen Paradox, from *Alice's Adventures in Wonderland*, is probably the appropriate alternative to the prisoners' dilemma story. Alice asks the red queen why she is running. The red queen answers that if she stopped, she would fall behind because she suspects that the others will keep on running. However, if the others likewise stopped running, she would not fall behind. The Red Queen differs from prisoners' dilemma game in one important respect. Prisoners' dilemma assumes that the prisoners are friends, i.e., have a commitment not to cheat each other. However, this begs the question: What is the origin of this commitment? In contrast, the Red Queen Paradox is not loaded with commitment [see Khalil, 1997a].

partnership (BP) if he expends low effort in the future. The agent gains the benefits of good partnership (GP) if he instead expends fair effort every day in the future.

		Future Self (infinite tomorrows)	
		Low Effort	Fair Effort
Current Self	Low Effort	10; 10,…; BP	10; 0,…; GP
	Fair Effort	0; 10, …; BP	0; 0, …; GP

Figure 3: The Fairness Paradox

As in the case of prudence, the dominant strategy of current self is to free-ride. That is, the agent is always better off to expend low effort in cleaning no matter what future self does.

If the agent applies this reasoning everyday, the agent ends up with loafing everyday, and hence, ends up with losing the benefit of partnership. This is undesirable outcome, assuming that the agent enjoys more the benefit of partnership than the pleasure of loafing throughout his lifetime.

What to do in order to avoid taking the rational dominant strategy? Similar to imposing precommitment to remedy imprudence, the agent may vote for a referee to remedy injustice. This is actually the core of the analysis of Armen Alchian and Harold Demsetz [1972]. They call the referee "monitor." The monitor imposes fines high enough to extinguish the 10 units of benefits from free-riding.[5]

4. A SIMPLE MODEL

To highlight the key assumptions, let us examine the game between "current self" and "future self." Current self (at time t=0) decides whether to indulge in an action (x) that deviates (d) from the decision which is optimum (x*). The action can be the consumption of food, alcoholic drinks, study, entertainment, work, sex, free-riding, loafing on the job, and so on. It is assumed that if x* is optimum for one period, it is optimum for T periods, where t= 0, …, T. If the agent over- or under-indulges repeatedly, wellness would be below the optimum depending on the extent and frequency

[5] Alchian and Demsetz explain the origin of referees and arbitration boards in disputes among business partners. But this does not mean, as they claim, to be the core of the employment contract that distinguishes the firm [Khalil, 1997b].

of the deviation. For instance, if one under-consumes food to an extreme degree and almost in all T periods, one can become anorexic. Likewise, if one over-consumes food extremely and frequently, one can become bulimic.

In this game, current self decides whether to deviate in order to maximum utility function (U)

$$U = U[\Sigma B^t U(x_t^*), \Sigma B^t E(d_t), W(\Sigma B^t d_t)] \quad \ldots \quad (1)$$

B^t is the discount factor. The first argument is the present value for acting according to the decision. The second argument denotes the present value of excitement (E) which is positive in d. The third argument denotes wellness (W) which is negative in deviation (d). In this manner, the opportunity cost of d is included in its impact on W. However, the out-of-pocket cost of d can be negative as in the case of anorexia or TV watching instead of exercising. But it is safe to assume that the out-of-pocket cost of d is zero since such cost is not the major variable in explaining weakness of will.

In this set up, it is assumed that

$$E(d_t) > |W(d_t)| \quad \ldots \quad (2)$$

that is, for any t, the excitement of the deviation from optimum exceeds the deterioration of wellness. However, for the present value of all future periods,

$$\Sigma B^t E(d_t) < |W(\Sigma B^t d_t)| \quad \ldots \quad (3)$$

That is, the present value of future excitements always gives lower utility than the absolute deterioration of wellness. The reason for this is simple. Wellness depends on a multiplicative factor of the deviations, while the total excitement is additive. This assumption corresponds to the common sense observation that, e.g., a single cigarette smoking would do miniscule or even zero impact on health, but would give a great pleasure. However, frequent cigarette smoking over one's life time would impact health greatly. Likewise, if one does not clean the apartment as agreed, it would not lead to the break up of cooperation. But repeated injustices over time plant the seed for resentment on the part of the injured party. The resentment might be suppressed for a while for the sake of friendship and cooperation. But eventually the resentment will erupt unexpectedly and lead to an abrupt break up of the partnership. Of course, the decline of wellness in both cases would vary depending on the spacing of smoking/loafing or, in general, the intensity and frequency of the deviations from the optimum. But this issue should not concern us here.

What matters here is that the weakness of will phenomenon is driven by a reasonable but potent assumption. Namely, the injuries done to the self by smoking, or to others by loafing, are multiplicative rather than additive. This feature sets the stage of why violation of the right of the self (imprudence) or the right of others (injustice) is the dominant strategy.

5. THREE THEORIES OF TRUST

Even though the violation of rights is the dominant strategy, why do agents still act in a trustworthy manner, i.e., respect the property rights of others and themselves?

There are three possible theories of trust—other than appealing to sociological norms which would beg the question [Khalil, 2003]. The first theory is the modelling of integrity as a strategy: The agent is simply afraid of the enormous retaliation, either by the monitor or by the self-imposed precommitment. This explanation actually denies the phenomenon it tries to explain: The agent cannot act in a trustworthy manner and, hence, the "police society" is necessary to monitor him in every step.

But from casual empiricism as mentioned earlier, there is no need for "police society." Agents do act in a trustworthy manner even when the dominant strategy is to cheat. So, the question is still on the table: Why is there trustworthiness?

The second theory is the modelling of integrity as a taste, i.e., as an element of the utility function. If so, agents with such tastes would be "stupid" in the market—i.e., act according to tastes that would put them at a disadvantage in the evolutionary game sense.

The third theory answers this objection. It models integrity as a character trait—what was called earlier "conscience"—as advanced, among many others, by Robert Frank [1987]. That is, integrity is a character trait that is favoured by societal selection. People prefer to deal with agents who have conscience or have the honesty trait. But this explanation begs the question, which faces all neo-Darwinian explanations: From where did the trait for honesty arise?

6. PROPRIETY

So, how to proceed? How to explain trust, i.e., the success of closing the decision-action gap without extensive precommitment or living in "police society"? How to explain the fact that agents do not often choose the dominant strategy of imprudence and injustice? How to explain the fact that weakness of will is not as widespread as the proposed model predicts?

While the current self is tempted to inflict injustice on others and imprudence on future self, current self often does not choose such actions. It is not the fear of punishment because, as shown, loafing is the dominant strategy.

The answer lies in the "propriety" mechanism advanced by Adam Smith [1976] in *The Theory of Moral Sentiments* [see Khalil, 1990]. The propriety mechanism highlights an alternative mechanism to punishment or the incentive provided by the external monitor. The propriety mechanism highlights the role of the internal monitor, what Smith called the "impartial spectator."

To understand propriety, we need to dissect the physiology of the self in Smith's theory of human conduct. The self seeks self-approval and, hence, enforces commitments via self-command. [6] To seek self-approval, according to Smith, current self is forced to moderate its appetites or emotions so that the distant self, i.e., the internal monitor, can enter and sympathize with the pleasure or pain of the current self. The current self wants the support and sympathy of the impartial spectator who is ultimately planted in the breast of the actor. The impartial spectator, who turns out to be the future self, looks at the pain and pleasures of the current self from a distance [Khalil, 1990]. Therefore, the impartial spectator can never feel the same intensity of emotions felt by the current self. And for the current self to get the approval of the impartial future self, it must lower the pitch of its emotions to match the view of the future self, i.e., the view from a distance. The impartial future self can travel and approve (i.e., sympathize) with the current self only if the current self moderates its emotions. By moderating the emotion, the current self effectively extinguishes the extra benefit (10 units in the above examples) it can achieve from loafing. So, the current self does not see the dominant strategy as appealing as would be the case if it is not interested in getting the approval of the impartial distant self.

But why should the current self seek the approval of the future, impartial self? The agent reaps pleasure from knowing that he has conquered the immediate appetites. At first approximation, there is no need for a police or external monitor. The self seeks to monitor its own impulses because self-command occasions the sense of self-respect or pride. For Smith, self-respect or pride can lead to vanity, self-aggrandizement, and ostentatious behaviour [Khalil, 1996]. However, once we look beyond the excesses, self-respect is a healthy motivation. So, agents do not only seek to maximize their welfare, they also seek to maximize self-respect, one of what I called elsewhere "symbolic product," afforded by self-command over immediate appetites and emotions [Khalil, 2000].

Thus, self-command or self-control over current excitements and appetites amounts to the negation of the dominant strategy. This mechanism of self-command is not necessarily an explicit,

---

[6] Thomas Schelling [1978, 1984a,b, 1992] also uses the term "self-command." But he actually uses it to denote "precommitment," i.e., punishment that locks one in a single, predictable action. So, Schelling's term "self-command" is identical to what Oliver Williamson [1983] calls "hostages": Firms may invest heavily in brand name advertisement or fancy building to inform suppliers and customers that they would not cheat.

calculative mechanism. Rather, it occurs instantaneously as the current self seeks sympathy and approval of its pain and pleasure from the impartial self.

Through the process self-approval, the agent attains integrity. The current action is integral of the grand optimizing plan of the agent as current self moderates the pitch of its excitement. Such a view of the complex self provides a mechanism to abridge the decision-action gap without the appeal to self-punishment or "police society."

Such a theory of the complex self opens new vistas of research. This cannot be elaborated here [Khalil, 2003]. But it suffices to emphasize that the view of the propriety mechanism as constructed daily, as the self attempts to maintain integrity, promises to supersedes the shortcomings of modelling self-integrity as a strategy to avoid punishment, as a taste, and as a character trait in evolutionary game.

7. IMPLICATIONS

The dominant view regards weakness of will an anomaly facing the standard theory of rationality. The paper argues the opposite: The fact that weakness of will is not widespread is the anomaly. In a simple model, the paper shows that weakness of will is the dominant strategy in a game between current self and future self. This leads to the motivating question of the paper: Why is weakness of will is not pervasive—given that precommitment and punishment are not sufficiently pervasive to remedy the weakness of will? The paper argues that the answer lies in what Adam Smith calls the "propriety" mechanism: Humans demand self-respect and, hence, exercise self-command over appetites and emotions.

The above analysis has many implications. First, is there a neural support to the hypothesis of propriety as the mechanism that abridges the decision-action gap? To wit, one neuroscientist, Michael Gazzaniga [1998, Gazzaniga *et al.*, 1998], argues that the brain has an operator, the "judge," that differs from the actor, the "doer." As a judge, the self is constantly evaluating the propriety of action. The integrated or healthy self primarily wants to make sure that the agent today is not hurting the agent tomorrow.

Second, the above analysis of self-command has been carried out with little reference to social interaction, the preferences of others, or even society. It is currently faddish among economists to introduce social interaction á la Gary Becker [1996] or even discuss "prosocial" preferences [Gintis *et al.,* 2005]. While social reference point influences the formation of self-command, the social group is neither a necessary nor a sufficient condition. One can model the agent as caring about his future self without invoking the preference of the social group.

Third, there is additional error when integrity or trust is lumped in the category of "prosocial" preferences. The fact that trust connotes some moral dimension, stemming from the preservation of integrity, does not mean automatically that it is a prosocial preference. One can think of altruism in the sense of charity as a prosocial preference. But trust or integrity is rooted in self-interest calculation, i.e., the protection of property either in the sense of justice or prudence. Trust has always been troubling to thinkers. On one hand, social contract theorists to David Hume [1751, 1896] and Richard Posner [1994] argue that property and trust is rooted in interest [see Khalil, 2005]. On the other hand, trust cannot be simply synonymous with interest—otherwise loafing and distrust would be widespread. Trust captures self-integrity. The simple model advanced here captures the double faces of trust: interest and integrity. The decision about property is rooted in interest, while the action to over-ride the dominant strategy of cheating expresses integrity.

Fourth, how could the agent close the decision-action gap if the decision is unknowable? The agent may not be too sure of whether his failure to execute a plan is the result of temptation or the result of over-evaluating his ability. If the agent over-evaluated his ability, and hence mistakenly decided on non-realistic goals the action may fall behind the decision for reasons other than weakness of will. This area of ambiguity is the source of much anxiety and turmoil, which is outside the scope of the paper.

Fifth, what is the relation between rationality and the emotions? Common wisdom for many centuries, as well as the wisdom of the man-on-the-street, is that one should control one's emotions for his own interest [Elster, 1999]. In fact, the opposition between reason and the emotions can be traced to Plato. However, a new line of thinking has flourished to demolish the dichotomy between reason and the emotions. As articulated by Jack Hirschleifer [1987] and Robert Frank [1988], natural selection has favoured agents who react emotionally in order to make credible threats against aggression: It is rational to be irrational and emotional so that aggressors do not think that the agent does not care about sunk cost. I call this argument "the invisible hand of the emotions." In this light, the set of emotions is the daughter of rationality: Even if natural selection did not select the over-reaction, the rational agent should demonstrate emotional over-reaction to scare others.

However, this argument entails that the emotions are socially dependent. One needs to live in society to have emotions and appetites. A Robinson Crusoe, who does not deal with any creature that understands his emotions, cannot employ emotions in the world of Hirschleifer and Frank. So, the emotions are society-dependent phenomena. But what about appetites related to the

emotions, such as craving for food when one is hungry, which are obviously not society-dependent phenomena?

To answer the question, we need to understand the emotions as appetites and sentiments which are not society-dependent. Adam Smith, in his *Theory of Moral Sentiments,* understood them as such. He did so remarkably without opposing reason and the emotions. The way Smith reconciled reason and the emotions is by modelling "reason" as simply the emotions or sympathy of the distant self, whom he called the "impartial spectator. So, for Smith, the emotions of the impartial spectator masquerade themselves as reason, while the emotions of the current self are the emotions proper.

Sixth, punishment, precommitment, and other forms of "police society" may crowd out propriety as the mechanism for abridging the decision-action gap. For instance, Richard Thaler and Cass Sunstein [2003] propose that libertarian paternalism, where the state solves the weakness of will with regard to saving, is not an oxymoronic concept. They might be correct. The point is rather whether the crowding out of propriety by precommitment can be expensive in the long run in terms of administrative cost. It might be more efficient to enact policies that nurture the growth of propriety—even though such policies may not have immediate results.

There are many other implications of the analysis of weakness of will. Weakness of will, as shown here, is not an anomaly in the temple of standard theory of rationality. What is anomalous in the temple is the absence of the widespread of weakness of will. This absence cannot be explained by the appeal to punishment, precommitment, or the "police society." One can only understand the absence by a careful physiology of the complex self. The physiology of the self should prove to be the next big challenge facing economics and social theory in general.

REFERENCES

Ainslie, George. *Breakdown of Will*. Cambridge; New York: Cambridge University Press, 2001.

Alchian, Armen A. and Harold Demsetz. "Production, Information Costs, and Economic Organization." *American Economic Review*, December 1972, 62:5, pp. 777-795.

Becker, Gary S. *Accounting for Tastes*. Cambridge, MA: Harvard University Press, 1996.

Elster, Jon. *Ulysses and the Sirens: Studies in Rationality and Irrationality*, revised edition. Cambridge: Cambridge University Press, 1984.

— *Strong feelings: Emotion, Addiction, and Human Behavior*. Cambridge, MA: MIT Press, 1999.

— *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge: Cambridge University Press, 2000.

Frank, Robert H. "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One With a Conscience?" *American Economic Review*, September 1987, 77:4, pp. 593-604.

— *Passions Within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton, 1988.

Gazzaniga, Michael S. *The Mind's Past*. Berkeley: University of California Press, 1998.

— Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience: The Biology of the Brain.* New York: W.W. Norton, 1998.

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr (eds.). *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life.* Cambridge, MA: MIT Press, 2005.

Hardin, Garrett James. "The Tragedy of the Commons." *Science*, 1968, 162, pp. 1243-1248.

Hirshleifer, Jack. "On the Emotions as Guarantors of Threats and Promises." In John Dupré (ed.) *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press, 1987, pp. 307-326.

Homer. *The Odyssey*, trans. by Ennis Rees. Indianapolis: Bobbs-Merrill, 1977.

Hume, David. *A Treatise of Human Nature*, 3 vols, ed. By L.A. Selby-Bigge. Oxford: Clarendon Press, [1740] 1896.

— *An Enquiry Concerning the Principles of Morals*. London: A. Millar, 1751.

Khalil, Elias L. "Beyond Self-Interest and Altruism: A Reconstruction of Adam Smith's Theory of Human Conduct." *Economics and Philosophy*, October 1990, 6:2, pp. 255-273.

— "Respect, Admiration, Aggrandizement: Adam Smith as Economic Psychologist." *Journal of Economic Psychology*, September 1996, 17:5, pp. 555-577.

— "The Red Queen Paradox: A Proper Name for a Popular Game." *Journal of Institutional and Theoretical Economics*, June 1997a, 153:2, pp. 411-415.

— "Is the Firm an Individual?" *Cambridge Journal of Economics*, July 1997b, 21:4, pp. 519-544.

— "Symbolic Products: Prestige, Pride and Identity Goods." *Theory and Decision*, August 2000, 49:1, pp. 53-77.

— "Why does Trustworthiness Pay? Three Answers: An Introduction." In Elias L. Khalil (ed.) *Trust*. Cheltenham, UK: Edward Elgar, 2003, pp. xiii-xxxii.

— "The Moral Justification of Enslavement." A working paper, 2005.

Montes, Leonidas. *Adam Smith in Context: A Critical Reassessment of Some Central Components of his Thought*. New York: Palgrave Macmillan, 2004.

Olson, Mancur. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press, 1965.

Posner, Richard A. "Law and Economics is Moral." In Robin Paul Malloy and Jerry Evensky (eds) *Adam Smith and the philosophy of law and economics*. Dordrecht: Kluwer, 1994, pp. 167-178.

Schelling, Thomas C. *The Strategy of Conflict*. London: Oxford University Press, 1960.

— "Egonomics, or the Art of Self-Management." *American Economic Review, Papers and Proceedings*, 1978, 68, pp. 290-294.

— "Self-Command in Practice, in Policy, and in a Theory of Rational Choice." *American Economic Review, Papers and Proceedings*, 1984a, 74, pp. 1-11.

— *Choice and Consequence*. Cambridge: Harvard University Press, 1984b.

— "Self-Command: A New Discipline." In George Loewenstein and Jon Elster (eds.), *Choice Over Time*. New York: Russell Sage Foundation, 1992, pp. 167-176.

Smith, Adam. *The Theory of Moral Sentiments*, edited by D.D. Raphael & A.L. Macfie. Oxford: Clarendon Press, 1976.

Thaler, Richard.H., and Cass R. Sunstein. "Libertarian Paternalism." *American Economic Review, Papers and Proceedings*, May 2003, 93:2, pp. 175-179.

Williamson, Oliver. "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, 73, pp. 519-540.