Ishida, Junichiro

Working Paper

Autonomy and motivation: A dual-self perspective

ISER Discussion Paper, No. 803

Provided in Cooperation with:
The Institute of Social and Economic Research (ISER), Osaka University

Suggested Citation: Ishida, Junichiro (2011) : Autonomy and motivation: A dual-self perspective, ISER Discussion Paper, No. 803, Osaka University, Institute of Social and Economic Research (ISER), Osaka

This Version is available at:
https://hdl.handle.net/10419/92892

Discussion Paper No. 803

AUTONOMY AND MOTIVATION: A DUAL-SELF PERSPECTIVE

Junichiro Ishida

February 2011

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Autonomy and Motivation: A Dual-Self Perspective*

Junichiro Ishida

ISER, Osaka University

February 15, 2011

Abstract

This paper provides a simple autonomy-based model of human motivation in which a decision maker with divided selves must perform some task. The key presumption of the model is that the brain is not a unitary system which is equipped to achieve a single goal in a systematic manner; rather, it is more like an organization which is hampered by several constraints such as preference incongruence and incomplete exchange (or imperfect recall) of information. Due to these constraints, the model yields behavioral patterns that are consistent with various stylized facts of human motivation, mostly found in social psychology. The main findings of the paper are: (i) more autonomy induces more motivation; (ii) complex tasks are susceptible to motivation crowding out; (iii) small rewards are detrimental to motivation; (iv) intrinsically interesting tasks are susceptible to motivation crowding out.

JEL Classification Codes: D03, D99.
Key Words: Autonomy, Intrinsic and extrinsic motivation, Dual self, Motivation crowding out.

*Correspondence to: Junichiro Ishida, Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan. E-mail address: jishida@iser.osaka-u.ac.jp.

1 Introduction

In economics, the term "incentive" is almost a synonym to motivation, and the distinction between them is rarely made clear. Economists typically emphasize the role of incentives in guiding and shaping human behavior without making much reference to motivation *per se*, as if they are simply two different ways of representing the same thing. This is probably due to the fact that economists are primarily concerned about extrinsic motivation – the side of human motivation which is induced by external enforcers, or "incentives," such as contingent rewards, monitoring, surveillance and evaluation.[1] To the extent that there exists a one-to-one monotonic relationship between motivation and incentives, the distinction between them is merely semantic and mixing them up does not lead to any substantial complications.

The problem is that human motivation is a far more complicated subject, and extrinsic motivation seems to represent only a part of it. It is now widely accepted, not only among social psychologists but also increasingly among economists, that to fully understand human motivation, one needs to understand the workings of intrinsic motivation, which arises from sheer pleasure of engaging in an activity itself, as well as those of extrinsic motivation. To further complicate the issue, abundance of evidence also suggests that these two concepts of motivation do not exist in isolation: there seem to exist substantial interactions between the two, where a change in one often leads to a change in the other. These findings imply that intrinsic and extrinsic motivation are not additive in nature, and focusing solely on extrinsic motivation, with no specific reference to intrinsic motivation, might lead to misleading predictions.

What we need is then a description of the underlying mechanism that can explain how intrinsic motivation interacts with extrinsic motivation. To this end, although there seems to exist no unified theory of human motivation to date, the recent literature in social psy-

[1] Of course, the distinction between motivation and incentive would ultimately come down to how we define them. A conventional view is probably that incentive is a means (to induce motivation) whereas motivation is an end. According to the Oxford Dictionary of English, for instance, it is defined as "a thing that motivates or encourages someone to do something." We take this view throughout the paper although we interchangeably use extrinsic rewards and incentives.

chology indicates several consistent patterns observed in a variety of settings, which potentially provide us with an angle to inquire into this issue. First, it is now well known and firmly established that external enforcers, such as rewards, surveillance, competition and evaluation, often undermine intrinsic motivation, thereby yielding a counterproductive effect (Deci, 1971; Deci and Ryan, 1985; Frey, 1997). Second, many studies also point out that autonomy, such as control over method or timing and discretion in setting goals, is a key determinant of intrinsic motivation (Hackman and Oldham, 1976; Gagne and Deci, 2005). Although this second fact is less known than the first, various sources argue that people tend to be more intrinsically motivated in doing an activity when it is their independent and "self-determined" choice to do so. Third, the saliency of intrinsic motivation depends on the type of task people do. In particular, intrinsic motivation seems to matter more for non-routine tasks, e.g., complex tasks or tasks that require creativity (Amabile et al., 1990; Oldham and Cummings, 1996). Forth, the hidden cost of incentives seems to be more salient when extrinsic rewards are relatively small in magnitude (Gneezy and Rustichini, 2000a). Finally, the hidden cost is more likely to surface when the target task is intrinsically interesting (Ryan et al., 1983; Deci et al., 1999).

In this paper, we construct a simple dual-self model of human motivation which is parsimonious and tractable enough to capture these observations. The key presumption of the model is that the brain is not a unitary system that is equipped to achieve a single goal in a cohesive manner; rather, it is like an organization which is hampered by various constraints such as preference incongruence and incomplete exchange (or imperfect recall) of information.[2] We consider a decision maker (DM) with divided selves, called Cognition and Affect, who must perform some task. Cognition is the sophisticated self who is endowed with cognitive ability to evaluate alternatives and make choice, whereas Affect is the primitive and instinctive self who is endowed with motivational or affective energy to carry out an action.[3] These differences in cognitive and affective characteristics give rise to the following

[2]One of the earliest works to model intrapersonal conflicts is Thaler and Shefrin (1981). Recently, there are increasingly many works which model a decision maker with divided selves. See Benabou and Tirole (2004), Bernheim and Rangel (2004), Fudenberg and Levine (2005) and Brocas and Carrillo (2008) for recent examples.

[3]This mean that Affect must take some initiative for one to produce an action. On this point, Camerer (2005) et al. note "our view is that cognition by itself cannot produce action; to influence behavior, the cognitive system

three features of the model.

- Decentralization/specialition: There are two stages in implementing an action: the planning stage (what to do) and the implementation stage (whether to do it). This sequence of choices is made in a decentralized way, where Cognition (mostly) does the first while Affect does the second.

- Incomplete information: The values, both extrinsic and intrinsic, of the action chosen in the first stage are not accessible to Affect who must infer them from the primitives of the environment.

- Preference incongruence: Cognition is more patient and forward-looking than Affect, because it requires a certain level of cognitive ability to expect and appreciate potential consequences of an action. This difference in time preferences effectively results in a conflict of interests between them, because the extrinsic and intrinsic values typically realize at different points in time.

To be more precise, the model goes as follows. In the first stage, either Cognition or Affect chooses an action to implement from a set of feasible alternatives, each of which is characterized by its intrinsic and extrinsic values. Initiative in this stage can be taken by either agent, depending on the complexity of the target task. In the second, Affect then decides whether to actually carry out the action chosen previously. The problem is that, due to the incomplete exchange or imperfect recall of information, it is not clear to Affect why that particular action was chosen in the first place and hence how rewarding the action is to him. Environmental factors, such as the degree of autonomy given to DM, rewards associated with the task, and the nature of the task, may then influence DM's motivation as represented by Affect's effort choice. Using this framework, we obtain conditions under which those environmental factors raise or undermine DM's motivation.

The paper is related to several strands of literature, both in social psychology and economics. The entire framework relies heavily on an insight developed in Self-Perception

must operate via the affective system."

Theory (Bem, 1972).[4] The theory posits that our attitudes and feelings are often uncertain and, when they are, we make an inference about these states by observing our past behaviors and the situation in which it occurs. To the extent that this holds true, external enforcers can inflict a negative impact on our intrinsic motivation. For instance, when people are paid to do an activity, they reason in retrospect that they do it because of the reward that they get. The unfortunate consequence is that when that happens, people often lose their interest in the activity which they initially do it for its own sake. In social psychology, this is often called the overjustification effect: extrinsic enforcers make them underestimate the intrinsic value of the activity. Alternatively, Cognitive Evaluation Theory (CET) argues that external enforcers diminish feelings of autonomy and undermine intrinsic motivation through a change in perceived locus of causality (PLOC) from internal to external. Not much is known, however, about the underlying mechanism that gives rise to the overjustification effect or a shift of PLOC. It is in particular not clear why we ever need to overjustify, rather than simply justify, our originating motives. Constructing a simple autonomy-based model, we provide a microfoundation for what appears to be overjustification.

Recently, the role of intrinsic motivation in general and the "hidden cost of rewards and incentives" in particular draw attention from economists as well and are discussed rather extensively (Kreps, 1997; Frey, 1997; Frey and Oberholzer-Gee, 1997; Gneezy and Rustichini, 2000a, 2000b; Fehr and List, 2004; Falk and Kosfeld, 2006).[5] Theoretical investigations of the hidden cost of incentives have also flourished. In a setting with an informed principal, Benabou and Tirole (2003) show that contingent rewards may be counterproductive since those rewards are taken as a signal about undesirable features of the task to be performed. Benabou and Tirole (2006) argue that rewards diminish the incentive for prosocial behavior because they undermine its signaling value. Sliwka (2007) considers a setting where some fraction of agents are conformists who derive utility from conforming to others. When those

[4]Hirshleifer and Welch (2002) provide a model with potential memory loss (actions are remembered well but information signals are not) which is related to Self-Perception Theory when it is brought into an individual decision problem. They then show when the decision maker exhibits excess inertia (insensitivity to new information) and excess impulsiveness (excess sensitivity to new information). Benabou and Tirole (2004) also consider a case where past attitudes are not certain and inferences must be made from past actions.

[5]Also see Frey (2001) for a survey.

agents do not know what fraction of agents are fair, trust emerges as a signal that most agents are indeed fair, and the principal may choose to trust, rather than control, agents. Ellingsen and Johannesson (2008) consider a similar setting to Benabou and Tirole (2006) but assume that the payoff of social esteem depends on the audience to whom the agent tires to impress. In that setting, they show that material incentives erode esteem incentives especially when the principal has a choice of incentive scheme.

All of these models mentioned above are concerned about interpersonal situations where things such as trust and esteem matter. In contrast, we focus on an intrapersonal mechanism of motivation crowding out as an alternative route. Since most experimental results are obtained in controlled laboratory environments, where interactions are mostly anonymous and social incentives are supposedly less renounced, we argue that the intrapersonal perspective can provide an insight that can complement the existing literature which focuses more on social incentives.[6]

The paper proceeds as follows. We outline the model in section 2 and analyze it in section 3. In section 4, we discuss main results of the model, especially when and under what conditions motivation crowding out occurs in this context. Finally, we offer some concluding remarks in section 5.

2 Model

2.1 Setup

Consider a two-stage model between agent C (cognition, she) and agent A (affect, he). These two selves altogether comprise a decision maker (DM) who must engage in some task, which we call the target task. The decisions are made in a hierarchical and sequential manner, where an action is chosen in the first (planning) stage and is implemented in the second (implementation) stage.

First stage: In the first stage, DM "finds" or "invents" an action to be implemented. The

[6]To clarify our stance in the paper, we do not mean to argue that interpersonal models cannot explain experimental results observed in laboratory settings. We do believe that social incentives work, either consciously or subconsciously, in those settings to some extent. See Ellingsen and Johannesson (2008) for a discussion on this issue. Our approach is hence not exclusive: there are many factors that come into play on this issue of human motivation, and the intrapersonal mechanism which we illustrate here possibly provides one such factor.

task in this context represents a (vague) goal that DM would like to achieve (such as "sell more" or "raise revenue") whereas the action represents a particular way to achieve it (how to proceed the task). Initiative in this stage can be taken by either agent A or agent C, depending on the complexity of the task. We assume that agent C is called upon to make this choice with probability $\rho \in [0,1]$ which measures the complexity of the target task: ρ is closer to one when the task is more complex and requires more cognitive resources; it is closer to zero when it is simpler and/or more routine.

Every possible action is completely characterized by its values (v,y), where v and y refer to the intrinsic and extrinsic values, respectively. The extrinsic value reflects a material benefit of the action and is often associated with an observable consequence of the action, e.g., the level of observable output, which takes some time to materialize. In contrast, the intrinsic value refers to other immediate gains that are intrinsic to the environment or to the action itself, such as the sense of achievement, responsibility, and job satisfaction which typically comes from inside DM.

While there are presumably many ways to model this process of "inventing" an action, we adopt an approach that is amenable to the standard consumption-choice framework. Suppose that each agent is endowed with some cognitive resource R which the agent in charge can allocate at his/her disposal. The agent must expend more resources to find an action with higher values, so that the cost of coming up with an action with (v,y) is simply given by $d(v+y)$ where $d > 0$ is some constant. The values can only be improved so that $(v,y) \in \mathbb{R}^2_+$. Define $r := R/d$ as the parameter capturing the degree of autonomy given to DM. The basic idea is that the larger r is, the larger the set from which DM chooses an action. In the most extreme case where DM is given no autonomy at all ($r = 0$), she has no choice but to implement the default action, possibly assigned by an outsider, whose values are normalized at $(0,0)$.

Second stage: After the action is chosen, agent A chooses the effort level $e \in \mathbb{R}_+$. The (motivational) cost of effort, which is incurred entirely by agent A, is given by $e^2/2$. What is critical here is that the values of the chosen action is not accessible to agent A, either because it is simply forgotten or because there is no established channel of communication between

the two selves.

2.2 Payoffs

The key to the entire analysis is the misalignment of preferences between the two selves. In particular, we would like to emphasize the difference in time preferences between them, which originates from the difference in cognitive ability to evaluate potential future consequences of an action. We assume that agent C is more patient and forward-looking than agent A because agent C can foresee future consequences better.[7] This difference in time horizon matters because each value materializes at different points in time. The intrinsic value tends to be realized more immediately than the extrinsic value because it stems mostly from sheer pleasure of performing the task itself. In contrast, the extrinsic value takes more time to materialize because it is contingent on the observable consequence of the action which may not be immediately available.

We specify each agent's preferences under this presumption. Given the values of the action and the effort level, the payoff for agent C is given by

$$U^{\mathrm{C}}(v, y, e) = u^{\mathrm{C}}(v, wy)e,$$

where $w \in \mathbb{R}_+$ is the incentive rate which may be imposed from outside. The incentive rate is the parameter of our utmost concern and should be interpreted broadly: it is meant to capture the salience of contingent rewards and punishments, surveillance, monitoring, evaluation and so on. The payoff for agent A is, on the other hand, given by

$$U^{\mathrm{A}}(v, y, e) = u^{\mathrm{A}}(v, wy)e - \frac{e^2}{2},$$

We make the following assumptions on the payoff functions.

Assumption 1 *The payoff functions $u^j : \mathbb{R}^2_+ \to \mathbb{R}_+$, $j = A, C$, are twice continuously differentiable and satisfy the following properties:*
(i) $u^j_1 > 0$, $u^j_2 > 0$, $u^j_{11} < 0$, $u^j_{22} < 0$;
(ii) $\lim_{v\to 0} u^j_1 = \infty$ *and* $\lim_{y\to 0} u^j_2 = \infty$.

[7] There is ample evidence that long-term and short-term rewards are processes in different regions of the brain. Also, Dohmen et al. (2010) show the correlation between cognitive ability and patience, implying that cognition plays some role in evaluating long-term rewards.

Assumption 2 *For any* (v, y) *and* w*,*

$$MRS^{\mathrm{A}} := \frac{u_1^{\mathrm{A}}}{u_2^{\mathrm{A}}} > MRS^{\mathrm{C}} := \frac{u_1^{\mathrm{C}}}{u_2^{\mathrm{C}}},$$

Assumption 2 is particularly important, which states that at any combination of the values, agent A is always more willing to trade the extrinsic value (a future benefit) for the intrinsic value (an immediate benefit) than agent C, capturing the idea that agent C is more patient and forward-looking.

3 The analysis

3.1 The second stage

Since agent A has no recollection of what has happened in the first stage, she must make an inference from the primitives of the model environment about the values of the chosen action. Let $(\hat{v}^j, \hat{y}^j)$, $j = A, C$, denote the expected choice of the values when the choice is made by agent j. Agent A's problem in the second stage is then defined as

$$\max_{e} \quad [\rho u^{\mathrm{A}}(\hat{v}^{\mathrm{C}}, w\hat{y}^{\mathrm{C}}) + (1-\rho)u^{\mathrm{A}}(\hat{v}^{\mathrm{A}}, w\hat{y}^{\mathrm{A}})]e - \frac{e^2}{2}.$$

The optimal effort, denoted by e^*, is given by

$$e^* = \hat{u}^{\mathrm{A}}(\rho) := \rho u^{\mathrm{A}}(\hat{v}^{\mathrm{C}}, w\hat{y}^{\mathrm{C}}) + (1-\rho)u^{\mathrm{A}}(\hat{v}^{\mathrm{A}}, w\hat{y}^{\mathrm{A}}).$$

Of course, in equilibrium, these expected choices must coincide with the actual choices.

3.2 The first stage

Let (v^j, y^j) denote the actual choice of the values when it is made by agent j. When agent j is called upon to make the choice, the agent solves

$$\max_{(v^j, y^j)} \quad u^j(v^j, wy^j)\hat{u}^{\mathrm{A}}(\rho),$$

subject to

$$v^j + y^j \leq r := \frac{R}{d}.$$

Since u is non-satiated, one can immediately see that the cognitive resource constraint always binds at the maximum. The first-order condition is given by

$$u_1^{\mathrm{j}} - w u_2^{\mathrm{j}} = 0, \tag{1}$$

i.e., the values are chosen so as to satisfy $MRS^j = w$ if there exists an interior solution. Note that under Assumption 1, an interior solution always exists, where the optimal solution can be written as $v^j(r,w)$ and $y^j(r,w)$.

Define $u^{jk} := u^j(v^k, wy^k)$, $u_1^{jk} := u_1^j(v^k, wy^k)$, $u_2^{jk} := u_2^j(v^k, wy^k)$ and so on. Then, (1) implies

$$\frac{u_1^{\mathrm{AC}}}{u_2^{\mathrm{AC}}} > w = \frac{u_1^{\mathrm{AA}}}{u_2^{\mathrm{AA}}}.$$

Figure 1 illustrates this for two extreme cases ($\rho = 0$ and $\rho = 1$).[8] Define

$$\mu(r,w) := \frac{u_1^{\mathrm{AC}}}{u_2^{\mathrm{AC}}} = \frac{u_1^{\mathrm{A}}(v^{\mathrm{C}}(r,w), wy^{\mathrm{C}}(r,w))}{u_2^{\mathrm{A}}(v^{\mathrm{C}}(r,w), wy^{\mathrm{C}}(r,w))},$$

which we use as a measure of the degree of preference incongruence between the two agents. The preferences are identical when $\mu(r,w) = w$ for any (r,w) and become more divergent as $\mu(r,w)$ becomes larger.

[Figure 1 about here]

4 What factors enhance (or diminish) motivation

The present framework allows us to examine how a change in the underlying environment influences DM's motivation which is directly represented by $e^* = \hat{u}^{\mathrm{A}}(\rho)$.

4.1 Autonomy

While the role of autonomy in inducing motivation is less discussed in economics, many theoretical hypotheses argue and corresponding experimental results suggest that autonomy plays a decisive role in determining the level of human motivation: people tend to be

[8] The first-stage problem is equivalent to $\max_{v,y} u^j(v,y)$ subject to $v + y/w \leq r$. The figures are illustrated this way, where the slope of the cognitive resource constraint is equal to $-w$.

more intrinsically motivated in doing an activity when it is their independent choice to do so. For instance, Cognitive Evaluation Theory (CET) argues that external enforcers diminish feelings of autonomy and undermine intrinsic motivation through a change in perceive locus of causality (PLOC) from internal to external. In contrast, giving choice about aspects of task engagement tends to enhance feelings of autonomy, prompting PLOC from external to internal. Similarly, Self-Determination Theory, a revised and broader version of Cognitive Evaluation Theory, clearly identifies autonomy as a driving force of human motivation.[9] Job Characteristic Theory also emphasizes autonomy as one of five important characteristics to enhance motivation (Hackman and Oldham, 1976). These theories imply that the allocation of authority and control right has some impacts on workers' motivation in a way that is not typically considered in economics.

We now examine how a change in the degree of autonomy r affects DM's motivation. To this end, it is convenient to establish the following notion.

Definition 1 *The intrinsic value is a normal good for agent C iff* $\partial v^C / \partial r > 0$.

It is widely accepted that most goods are indeed normal, implying that the condition should not be particularly restrictive: in fact, this property holds for a wide class of preferences, such as CES, typically used in economic analyses. When this condition is satisfied, we can establish the following result.

Proposition 1 *An increase in r raises DM's motivation (the equilibrium effort level* e^**) if the intrinsic value is a normal good.*

PROOF: See Appendix.

The intuition behind this result should be clear and obvious. It simply says that agent A, who has no recollection of what has happened in the first stage, can reason that an action which is chosen from a wider set of alternatives is more likely to be of some value. In contrast, when DM is given less autonomy, an action is expected to be imposed more from

[9] See, for instance, Gagne and Deci (2005) for a survey.

outside and there is little reason to believe that it accords well with agent A's preferences. Figure 2 illustrates this when $\rho = 0$. In the figure, an increase in r raises Agent C's payoff level from $\bar{u}_0^C$ to $\bar{u}_1^C$ and moves the equilibrium point accordingly. One can see that this change induces more motivation from agent A as her payoff level improves at this new equilibrium point.

[Figure 2 about here]

4.2 Extrinsic rewards

We now turn to the main focus of the analysis, i.e., the effect of a change in extrinsic rewards on motivation. For expositional purposes, we say that motivation crowding out occurs if an increase in w reduces the equilibrium effort level e^* or, equivalently, agent A's expected payoff. There is now abundance of evidence which documents the undermining effect of extrinsic rewards (the case of motivation crowding out), both in social psychology and economics. The previous findings are so overwhelming that the relevant question to be asked is not whether this undermining effect exists, but when and under what conditions the effect is more likely to surface. In what follows, we raise several factors and discuss them in turn.

4.2.1 Complex tasks are susceptible to motivation crowding out

Cognition plays a larger role in choosing what to do when the task at hand is complicated and/or requires creativity in some sense. A number of observations in fact suggest that intrinsic motivation matters more for those complex, non-routine, tasks. First, many argue that high intrinsic motivation is a necessary ingredient for achieving creativity (Amabile, 1996; Shalley and Oldham, 1997). Moreover, it is also suggested that the complexity of tasks and intrinsic motivation (and creativity) are closely connected (Amabile, 1996; Boomer and Jalajas, 2002; Hackman and Oldham, 1980). Given this nature, it is understandable that motivation crowding out is more salient for complex tasks or tasks that require creativity. Amabile et al. (1990) show that external factors such as evaluation and competition can be detrimental to creativity. McGraw and McCullers (1979) find that monetary rewards also diminish cognitive flexibility in problem solving while Erez et al. (1990) find that monetary

rewards diminish performance on a complex task with difficult goals. Of course, complex tasks and tasks that require creativity are not always the same. Baer et al. (2003) show that extrinsic rewards lower creativity for employees in complex tasks. All of these studies indicate that extrinsic rewards make people think that they perform the task for the rewards that they get, thereby inhibiting intrinsic motivation that is necessary for enhancing performance in complex tasks.

To examine the undermining effect of extrinsic rewards, it is convenient to use the following notion.

Definition 2 *The intrinsic and extrinsic values are gross substitutes for agent C if* $\partial v^C / \partial w < 0$.

Two goods are said to be gross substitutes if the Marshallian demand for one good increases when the price of the other good increases.[10] Again, the condition for gross substitutability is not particularly restrictive: in the case of preferences represented by CES, the goods are gross substitutes if the elasticity of substitution is greater than one.

In the present context, the complexity of the task is measured by ρ, where a larger ρ means that the target task is more complicated and hence agent C is more likely to take over the decision-making process. We then obtain the following result which shows that motivation crowding out is more likely for complex, non-routine, tasks.

Proposition 2 *Suppose that (i) the values are gross substitutes and (ii) the preferences are sufficiently divergent, i.e.,* $\mu(r, w)$ *is sufficiently large. Then, motivation crowding out occurs if* ρ *is sufficiently close to one.*

PROOF: See Appendix.

[Figure 3 about here]

The proposition indicates that motivation crowding out is more likely to occur when the preferences are divergent (a large μ) and the target task is relatively complex (a large

[10] Since the price of the extrinsic value can be regarded as $1/w$, the two values are gross substitutes for agent C if $\partial v^C / \partial w < 0$.

ρ). These conditions are closely related and complementary to each other. When the task is simple and the first-stage choice is made by agent A, there is little room for preference incongruence to play any role, meaning that motivation crowding out is less likely to occur. In contrast, as the task becomes more complicated, an increase in w sways the choice excessively towards the extrinsic value (overjustification), from the viewpoint of Agent A, thereby lowering her subsequent payoff. The following result establishes necessary conditions for motivation crowding out in the current context.

Proposition 3 *Motivation crowding out never occurs if* $\rho = 0$ *or* $\mu(r, w) = w$ *for all* (r, w).

PROOF: See Appendix.

This result shows that preference incongruence within oneself is the driving force of motivation crowding out, which provides an answer to the question we set out at the beginning, i.e., why we ever need to overjustify, rather than justify, our originating motives. To see this, note that our model is the one of perfectly rational agents where agent A rationally forms expectations about agent C's behavior: a change in w changes agent C's choice of the values, but any change in the choice is always correctly anticipated by agent A. Our model shows that what appears to be overjustification can be understood within the framework which strictly rests on rational behavior, without relying on irrational overjustification, as long as there is some degree of preference incongruence within one self. In contrast, one can also argue that the presence of preference incongruence is the necessary condition for motivation crowding out in the current setup.

4.2.2 Small rewards are detrimental to motivation

Some of recent experiments show that extrinsic rewards undermine intrinsic motivation especially when the rewards are small in magnitude, i.e., no rewards are better than small rewards (Gneezy and Rustichini, 2000a). Intuitively, when no rewards are offered at all, the activity must be done purely for intrinsic motivation. Small rewards are then detrimental because they shift agent C's attention more towards the extrinsic value, which makes pref-

erence incongruence more resounding.

Proposition 4 *Suppose that the values are gross substitutes and* $\lim_{y\to 0} yu_2^{\mathrm{A}} = 0$. *Then, motivation crowding out occurs for any* $\rho \in (0,1]$ *if* w *is sufficiently small.*

PROOF: See Appendix.

[Figure 3 about here]

This result stems from the fact that motivation crowding out is more likely when y^{A} and y^{C} are relatively small. To see this, we need to consider both the direct effect and the indirect effect of changes in the incentive rate w. First, fixing the choice of the values, a unit increase in w directly raise agent A's payoff by $\rho y^{\mathrm{C}} u_2^{\mathrm{AC}} + (1-\rho) y^{\mathrm{A}} u_2^{\mathrm{AA}}$, which is positive by design. Second, a change in w shifts the choice of the values and indirectly affects his payoff. The second, indirect, effect is negative when the values are gross substitutes, and motivation crowding out surfaces when the indirect effect dominates the direct effect. It is then clear that when y^{A} and y^{C} are smaller, the direct effect is weaker and more likely to be dominated by the indirect effect. Small rewards are then detrimental to motivation because the marginal payoff from the extrinsic value is low when the incentive rate w is low, making the direct effect less of a factor.

4.2.3 Intrinsically interesting tasks are susceptible to motivation crowding out

A consensus among social psychologists is that the target task must be intrinsically interesting to observe any effects on intrinsic motivation. Due to this widely held perception, most experiments in fact examine the effect of external rewards for interesting activities (Ryan et al., 1983; Deci et al., 1999). To examine this issue, we define

$$\phi(r,w) := \frac{y^{\mathrm{C}}(r,w)}{w v^{\mathrm{C}}(r,w)},$$

and $\phi_{\max} := \max_{(r,w)} \phi(r,w)$. When the marginal return to raising y is small, agent C expends more resources to raise v. We thus take $\phi_{\max}$ as measuring how intrinsically interest-

ing the target task is: we say that the task is intrinsically interesting and challenging when $\phi_{\max}$ is small.

Proposition 5 *Suppose that the values are gross substitutes. Then, motivation crowding out occurs for any $\rho \in (0, 1]$ if $\phi_{\max}$ is sufficiently small.*

PROOF: See Appendix.

5 Conclusion

This paper provides a simple dual-self model of human motivation which allows us to inquire into the effects of various environmental factors on human motivation. The present model shows that extrinsic rewards may indeed produce a counterproductive effect and clarify under what conditions this negative consequence of rewards is more likely to surface. The main findings of the paper are: (i) more autonomy induces more motivation; (ii) complex tasks are susceptible to motivation crowding out; (iii) small rewards are detrimental to motivation; (iv) intrinsically interesting tasks are susceptible to motivation crowding out.

A natural next step is to take the present framework into interpersonal contexts. The incentive rate w and the degree of autonomy r are noth likely to be under the principal's discretion. If the principal cares only about the extrinsic value, e.g., the output, then she faces an interesting tradeoff. Giving more autonomy to the agent results in the action that is fun (a high intrinsic value) but not so productive. The cost arising from this distorted choice of action may, however, be more than compensated by an increase in the agent's motivation. This in fact seems to be the tradeoff faced by many managers: the tradeoff between less control and more motivation. The present framework can be extended to shed light on this aspect, and it is of some importance to address this issue in future.

Appendix: proofs

PROOF OF PROPOSITION 1: An increase in r raises DM's motivation iff

$$\rho\Big(u_1^{\mathrm{AC}}\frac{\partial v^{\mathrm{C}}}{\partial r}+wu_2^{\mathrm{AC}}\frac{\partial y^{\mathrm{C}}}{\partial r}\Big)+(1-\rho)\Big(u_1^{\mathrm{AA}}\frac{\partial v^{\mathrm{A}}}{\partial r}+wu_2^{\mathrm{AA}}\frac{\partial y^{\mathrm{A}}}{\partial r}\Big)>0,$$

$$\frac{\partial v^j}{\partial r}+\frac{\partial y^j}{\partial r}=1,\ j=A,C.$$

This can be written as

$$\rho\Big((u_1^{\mathrm{AC}}-wu_2^{\mathrm{AC}})\frac{\partial v^{\mathrm{C}}}{\partial r}+wu_2^{\mathrm{AC}}\Big)+(1-\rho)wu_2^{\mathrm{AA}}>0.$$

Since we already know that $u_1^{\mathrm{AC}}>wu^{\mathrm{AC}}$, this condition holds if $\partial v^{\mathrm{C}}/\partial r>0$, i.e., the intrinsic value is a normal good.

Q.E.D.

PROOF OF PROPOSITION 2: An increase in w lowers DM's motivation iff

$$\rho\Big(u_1^{\mathrm{AC}}\frac{\partial v^{\mathrm{C}}}{\partial w}+wu_2^{\mathrm{AC}}\frac{\partial y^{\mathrm{C}}}{\partial w}+y^{\mathrm{C}}u_2^{\mathrm{AC}}\Big)+(1-\rho)\Big(u_1^{\mathrm{AA}}\frac{\partial v^{\mathrm{A}}}{\partial w}+wu_2^{\mathrm{AA}}\frac{\partial y^{\mathrm{A}}}{\partial w}+y^{\mathrm{A}}u_2^{\mathrm{AA}}\Big)<0,$$

where

$$\frac{\partial v^j}{\partial w}+\frac{\partial y^j}{\partial w}=0,\ j=A,C.$$

It follows from these that

$$\rho\Big((u_1^{\mathrm{AC}}-wu_2^{\mathrm{AC}})\frac{\partial v^{\mathrm{C}}}{\partial w}+y^{\mathrm{C}}u_2^{\mathrm{AC}}\Big)+(1-\rho)y^{\mathrm{A}}u_2^{\mathrm{AA}}<0, \tag{2}$$

As $\rho\to 1$, (2) becomes

$$(u_1^{\mathrm{AC}}-wu_2^{\mathrm{AC}})\frac{\partial v^{\mathrm{C}}}{\partial w}+y^{\mathrm{C}}u_2^{\mathrm{AC}}<0,$$

which is further reduced to

$$(\mu(r,w)-w)\frac{\partial v^{\mathrm{C}}}{\partial w}+y^{\mathrm{C}}<0.$$

Given that the values are gross substitutes, this holds if $\mu(r,w)$ is sufficiently large.

Q.E.D.

PROOF OF PROPOSITION 3: When $\rho = 0$, (2) is reduced to $y^{\mathrm{A}} u_2^{\mathrm{AA}} > 0$, so that this condition never holds. When $\mu(r, w) = w$ for all (r, w), on the other hand, (2) becomes

$$\rho y^{\mathrm{C}} + (1 - \rho) y^{\mathrm{A}} \frac{u_2^{\mathrm{AA}}}{u_2^{\mathrm{AC}}} < 0,$$

which never holds either.

Q.E.D.

PROOF OF PROPOSITION 4: Note that as $w \to 0$, $y^j \to 0$. If $\lim_{y \to 0} y u_2^{\mathrm{A}} = 0$, (2) becomes

$$(u_1^{\mathrm{AC}} - w u_2^{\mathrm{AC}}) \frac{\partial v^{\mathrm{C}}}{\partial w} < 0,$$

for any $\rho > 0$. It is immediate to see that this condition holds if the values are gross substitutes.

Q.E.D.

PROOF OF PROPOSITION 5: Motivation crowding out occurs if

$$\rho \Big((\mu(r, w) - w) \frac{\partial v^{\mathrm{C}}}{\partial w} + y^{\mathrm{C}} \Big) + (1 - \rho) y^{\mathrm{A}} \frac{u_2^{\mathrm{AA}}}{u_2^{\mathrm{AC}}} < 0. \tag{3}$$

Dividing both side of (3) by $w v^{\mathrm{C}}(r, w)$ we obtain

$$\rho \Big(\frac{\mu(r, w) - w}{w v^{\mathrm{C}}(r, w)} \frac{\partial v^{\mathrm{C}}}{\partial w} + \phi(r, w) \Big) + (1 - \rho) \frac{y^{\mathrm{A}}}{w v^{\mathrm{C}}(r, w)} \frac{u_2^{\mathrm{AA}}}{u_2^{\mathrm{AC}}} < 0.$$

Since $y^{\mathrm{C}}(r, w) > y^{\mathrm{A}}(r, w)$ for any (r, w), it suffices to show that

$$\rho \Big(\frac{\mu(r, w) - w}{w v^{\mathrm{C}}(r, w)} \frac{\partial v^{\mathrm{C}}}{\partial w} + \phi(r, w) \Big) + (1 - \rho) \phi(r, w) \frac{u_2^{\mathrm{AA}}}{u_2^{\mathrm{AC}}} < 0.$$

As $\phi_{\max} \to 0$, the condition becomes

$$(\mu(r, w) - w) \frac{\partial v^{\mathrm{C}}}{\partial w} < 0,$$

which holds if the values are gross substitutes.

Q.E.D.

References

Amabile, Teresa M., 1996, *Creativity in Context*, Boulder, CO: Westview Press.

Amabile, Teresa M.; Goldfarb, Phyllis and Brackfield, Shereen C., 1990, Social Influences on Creativity: Evaluation, Coaction, and Surveillance, *Creativity Research Journal*, 2, 231-53.

Baer, Markus; Oldham, Greg R. and Cummings, Anne, 2003, Rewarding Creativity: When Does It Really Matter? *Leadership Quarterly*, 14, 569-86.

Bem, Daryl J., 1972, Self-Perception Theory, In Berkowitz, L. (Ed.), *Advances in Experimental Social Psychology*, New York: Academic.

Benabou, Roland and Tirole, Jean, 2003, Intrinsic and Extrinsic Motivation, *Review of Economic Studies*, 62, 315-39.

Benabou, Roland and Tirole, Jean, 2004, Willpower and Personal Rules, *Journal of Political Economy*, 112, 848-86.

Benabou, Roland and Tirole, Jean, 2006, Incentives and Prosocial Behavior, *American Economic Review*, 96, 1652-78.

Bernheim, B. Douglas and Rangel, Antonio, 2004, Addiction and Cue-Triggered Decision Processes, *American Economic Review*, 94, 1558-90.

Boomer, Michael and Jalajas, David, 2002, The Innovation Work Environment of High-Texh SMEs in the USA and Canada, *R&D management*, 32, 379-86.

Brocas, Isabelle and Carrillo, Juan D., 2008, The Brain as a Hierarchical Organization, *American Economic Review*, 98, 1312-46.

Camerer, Colin; Loewenstein, George and Prelec, Drazen, 2005, Neuroeconomics: How Neuroscience Can Inform Economics, *Journal of Economic Literature*, 43, 9-64.

Deci, Edward L., 1971, Effects of Externally Mediated Rewards on Intrinsic Motivation, *Organizational Behavior and Human Performance*, 8, 217-29.

Deci, Edward L. and Ryan, Richard M., 1985, *Intrinsic Motivation and Self-Determination in Human Behavior*, New York: Plenum.

Deci, Edward L.; Koestner, Richard and Ryan, Richard M., 1999, A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation, *Psychological Bulletin*, 125, 627-68.

Dohmen, Thomas; Falk, Armin; Huffman, David, and Sunde, Uwe, 2010, Are Risk Aversion and Impatience Related to Cognitive Ability?, *American Economic Review*, 100, 1238-60.

Ellingsen, Tore and Johannesson, Magnus, 2008, Pride and Prejudice: The Human Side of Incentive Theory, *American Economic Review*, 98, 990-1008.

Erez, Miriam; Gopher Daniel and Arzi Nira, 1990, Effects of Goal Difficulty, Self-Set Goals, and Monetary Rewards on Dual Task Performance, *Organizational Behavior and Human Decision Processes*, 47, 247-69.

Falk, Armin and Kosfeld, Michael, 2006, The Hidden Costs of Control, *American Economic Review*, 96, 1611-30.

Fehr, Ernst and List, John A., 2004, The Hidden Costs and Returns of Incentives: Trsut and Trustworthiness among CEOs, *Journal of the European Economic Association*, 2, 743-71.

Frey, Bruno S., 1997, *Not Just for The Money: An Economic Theory of Personal Motivation*, Cheltenham, UK and Brookfield, USA: Edward Elgar.

Frey, Bruno S. and Jegen Reto, 2001, Motivation Crowding Theory, *Journal of Economic Surveys*, 15, 589-611.

Frey, Bruno S. and Oberholzer-Gee, Felix, 1997, The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding Out, *American Economic Review*, 87, 746-55.

Fudenberg, Drew and Levine, David K., 2006, A Dual-Self Model of Impulse Control, *American Economic Review*, 96, 1449-76.

Gagne, Marylene and Deci, Edward L., 2005, Self-Determination Theory and Work Motivaion, *Journal of Organizational Behavior*, 26, 331-62.

Gneezy, Uri and Rustichini, Aldo, 2000a, Pay Enough or Don't Pay At All, *Quarterly Journal of Economics*, 115, 791-810.

Gneezy, Uri and Rustichini, Aldo, 2000b, A Fine is a Price, *Journal of Legal Studies*, 29, 1-18.

Hackman, J. Richard and Oldham, Greg R., 1976, Motivation through the Desing of Work, *Organizational Behavior and Human Performance*, 16, 250-79.

Hackman, J. Richard and Oldham, Greg R., 1980, *Work Redesign*, Reading, MA: Addison-Wesley.

Hirshleifer, David and Welch, Ivo, 2002, An Economic Approach to the Psychology of Change: Amnesia, Inertia, and Impulsiveness, *Journal of Economics and Management Strategy*, 11, 379-421.

Kreps, David M., 1997, Intrinsic Motivation and Extrinsic Incentives, *American Economic Review*, 87, 359-64.

McGraw, Kenneth O. and McCullers, John C., 1979, Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set, *Journal of Experimental Social Psychology*, 15, 285-94.

Oldham, Greg R. and Cummings, Anne, 1996, Employee Creativity: Personal and Contextual Factors at Work, *Academy of Management Journal*, 39, 607-34.

Ryan, Richard M.; Mims, Valerie and Koestner, Richard, 1983, Relation of Reward Contingency and Interpersonal Context to Intrinsic Motivation: a Review and Test Using Cognitive Evaluation Theory, *Journal of Personality and Social Psychology*, 45, 736-50.

Shalley, Christina E. and Oldham, Greg R., 1997, Competition and Creative Performance: Effects of Competitor Presence and Visibility, *Creative Research Journal*, 10, 337-45.

Sliwka, Dirk, 2007, Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes, *American Economic Review*, 97, 999-1012.

Thaler, Richard H. and Shefrin, H.M., 1981, An Economic Theory of Self-Control, *Journal of Political Economy*, 89, 392-406.

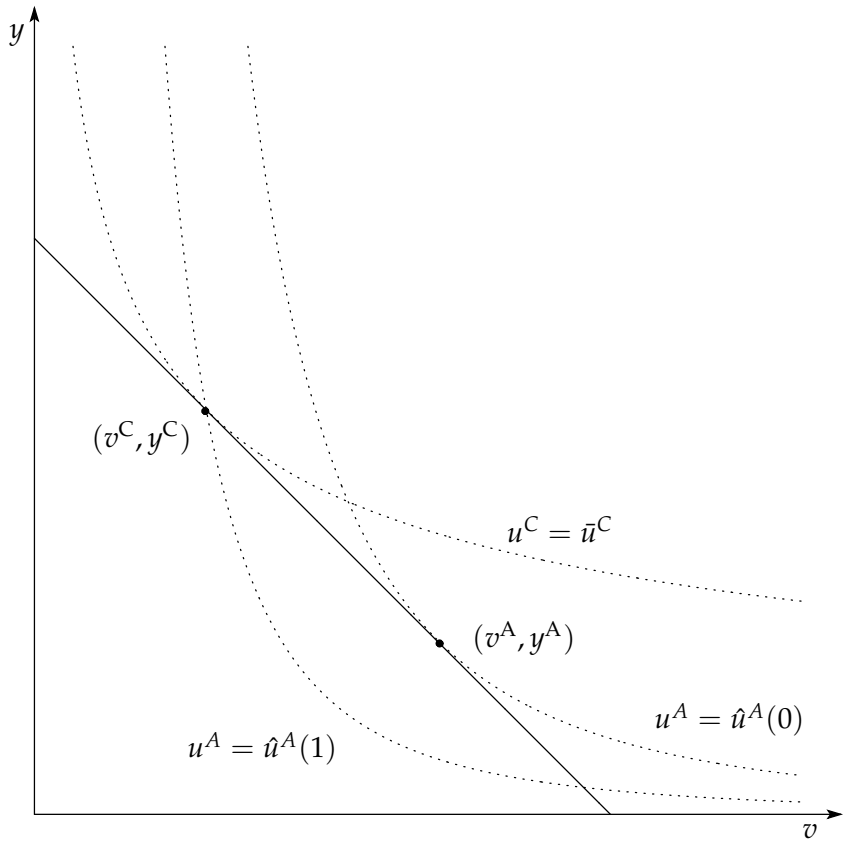

Figure 1: The equilibrium choice of the values and the corresponding motivation level.

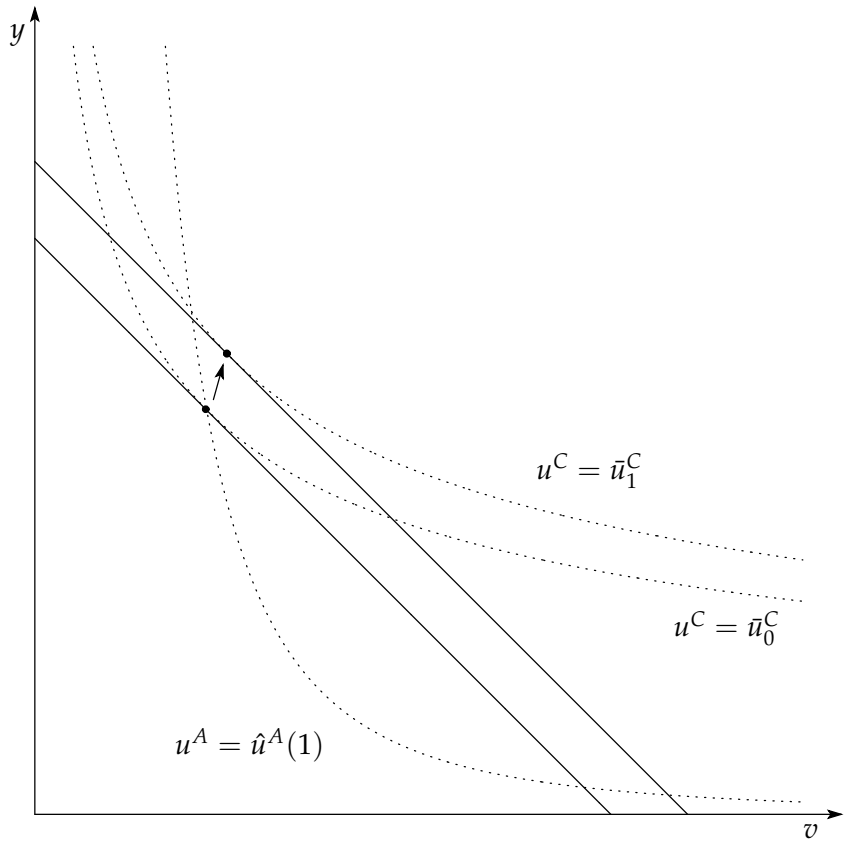

Figure 2: Autonomy enhances motivation ($\rho = 1$).

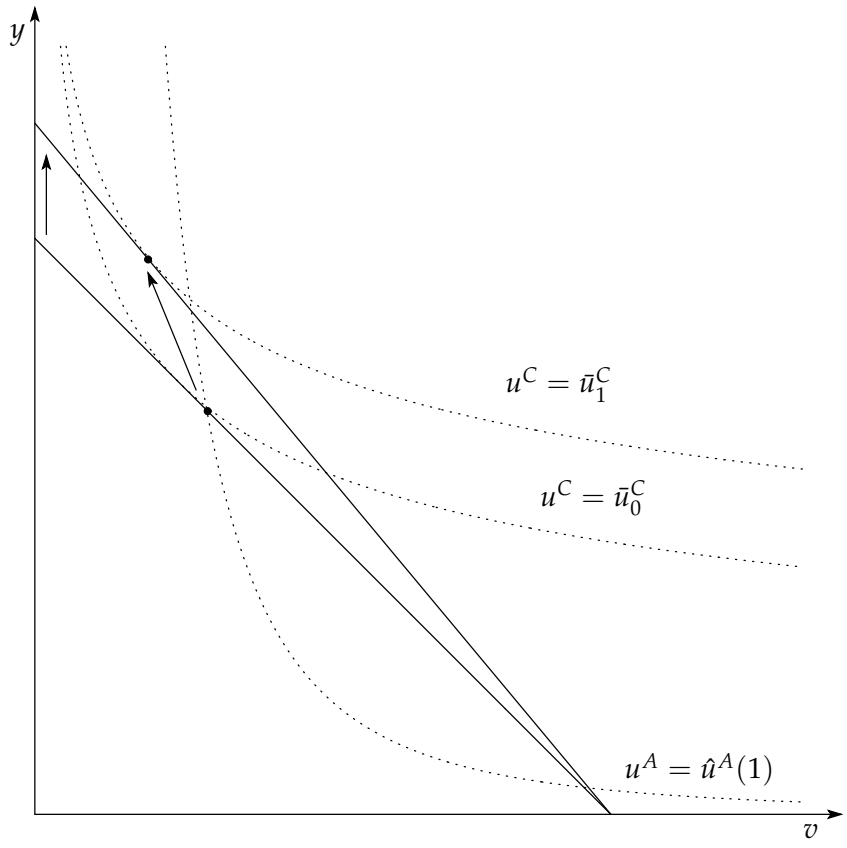

Figure 3: The case of motivation crowding out ($\rho = 1$).