THE INTERACTION BETWEEN NORMS AND ECONOMIC INCENTIVES[†]

Intrinsic Motivation and Extrinsic Incentives

By DAVID M. KREPS*

According to social psychologists and sociologists, a norm is a somewhat general rule of voluntary behavior. Examples range from the very general norm of reciprocity—treat others as they treat you—to more specific rules such as tip 15 percent and face the front in a crowded elevator.

Why do people adhere to norms? Economists have available four answers: (i) Adherence is costless relative to violation and so, why not? (ii) Adherence is immediately personally beneficial because it permits coordination (e.g., bear to the right in a crowded walkway). (iii) Adherence, while immediately costly, leads to better treatment by others than will violation. (iv) Adherence is desirable per se.

I reject (i) as uninteresting. Norms to achieve coordination, also called customs or focal points, are interesting, but I will not discuss them further here. Explanation (iii) spins out into the usual game-theoretic story that employs folk-theorem/reputation constructions for repeated games. Predictable commentary about observability being crucial, noise being inimical, and so on may be taken as read. As for (iv), since choice/utility theory is based on revealed preference, this involves making adherence (either to norms in general or to this specific norm) an argument in the individual's utility function.

When norms and economic incentives interact, the distinction between explanations (iii) and (iv) can be important. For example, Assar Lindbeck et al. (1996) study how the political economy of welfare is affected by the social norm that individuals should earn their own bread. In their model, adherence to this norm enters the utility function directly [they use explanation (iv)] but their analysis would not change much if they followed instead the lead of Douglas Bernheim (1994), who uses the desire to obtain social esteem [a type-(iii) rationale] as the basis for adherence. Imagine, however, that welfare payments can be hidden (in specific cases) from the general public. (Imagine combining workfare and earned-income tax credits.) Such "opacifying" institutional features may be desirable, because they shield the truly unfortunate from social stigma. But how do they affect the welfare rolls? If the good opinion of others keeps people off the "dole," then opacifying welfare administration leads to a large (equilibrium) increase in people on the dole. If the norm works through pure "self-respect," however, the number of workers relying on welfare would not change. Or, in the context of norms in the workplace, job design can affect the transparency of one worker's actions to others. If beneficial workplace norms work via (iii), the desirability of transparency is increased. If they work via (iv), increased opacity might be chosen to serve other objectives (e.g., to promote the privacy rights of individuals).

Do people adhere to the norm because it is a norm per se, or is there something desirable in the specific norm? This may be important to resolve, for example, if one wishes to suppose that the power of a norm (to command adherence) is positively associated with its level of adherence in the population (Lindbeck et al., 1996). If explanation (iii) is accepted for adherence, it is nearly tautological that

[†] *Discussants:* Oliver Williamson, University of California-Berkeley; Glenn Loury, Boston University.

* Graduate School of Business, Stanford University, Stanford, CA 94305, and Berglas School of Economics, Tel Aviv University, Israel. I am grateful to James Baron, Michael Hannan, Bengt Holmstrom, Eddie Lazear, Glenn Loury, John McMillan, Michael Morris, Joel Podolny, Paul Romer, and Oliver Williamson for helpful discussion and comments. The financial assistance of the National Science Foundation (SBR-9511208) is gratefully acknowledged.

norms work because they are norms; adherence relies entirely on their general acceptance and thus enforcement by the population. But if adherence directly enters the individual's utility function, there can be both models in which some behavior cannot be norm-induced and alternatives where adherence to norms per se enters the utility function, and thus any behavior can be norm-induced. Suppose the Departments of Economics at Harvard and the University of Chicago have different norms regarding the treatment of junior colleagues. Suppose eminent Professor X moves from Chicago to Harvard. Can Professor X be expected to adhere fully to Harvard's norms, upon arrival? If (iii) drives adherence, or if (iv) is the basis and conforming per se is valued, then the answer is yes. But if (iv) drives adherence, and it is specific behavior patterns that are internalized, then the answer is "not so readily." While transfers between Chicago and Harvard are too rare and unimportant to be of consequence, issues of labor mobility or managing multinationals or merged enterprises depend on the answer to this question.

Important aspects of the interaction between norms and economic incentives in organizations turn on the answers to these two and similarly murky questions. I do not know the answers, and an unscientific survey of colleagues from social psychology and sociology leaves me believing that they are not settled empirical matters. Instead, I am left believing that the answers change depending on circumstances. A useful general theory of interactions is impossible for now (at least). Instead, I will try to build some pre-empirical intuition by considering a single interesting case.

I. A Stylized Fact

The Stanford Graduate School of Business course in Human Resource Management is jointly taught by economists, sociologists, and social psychologists. As part of the social contract, my noneconomist colleagues use terms such as transaction costs, governance, agency theory, and the folk theorem, and I must assert things that, as an economist, I do not really understand. For example, I assert that providing extrinsic incentives for workers can be counterproductive, because it may destroy the workers' intrinsic motivation, leading to lessened levels of quality-weighted effort and lower net profits for the employer. This is not to say that intrinsic motivation is always superior to extrinsic incentives. In well-documented cases, the imposition of extrinsic incentives can lead to significant increases in worker effort and employer profit (Edward Lazear, 1996). The stylized fact applies only (primarily?) to employees with high initial levels of intrinsic motivation, where pride in one's work is high and the work is interesting.

Even with this caveat, strong empirical support of the stylized fact is hard to find. Classic studies include Mark Lepper et al. (1973), who document the effect on nursery school children, and E. Deci (1971), who gives experimental results suggesting this effect. But these and subsequent studies can be given other interpretations (Barry Staw, 1989).

On the other hand, anecdotal evidence abounds. And from the perspective of this session, this "fact" is very interesting. How can increased economic incentives lead to a diminution in worker effort and make the principal worse off? Understanding intrinsic motivation and this asserted interaction with economic incentives seems tailor-made for this session. So while there may be nothing to explain here, I will assume there is something to the stylized fact; abundant smoke signifies a fire, and the assertion is too strongly rooted in folk wisdom to be entirely hot air.

II. Intrinsic Motivation?

In the standard (simple) model of agency theory, introducing extrinsic incentives cannot lower effort levels; without extrinsic incentives, effort is necessarily at the lowest possible level. To explain the stylized fact, one first must answer: What is intrinsic motivation? Without extrinsic incentives, why would a worker expend any effort?

It is hard to imagine an employment situation without any extrinsic incentives whatsoever. In most employment situations where intrinsic motivation is meant to be high, the employee usually desires continued employment: he forms personal associations with coworkers, and he develops capital specific to his particular job and employer. Involuntary dis-

charge means uncertainty about future work and, perhaps, costly relocation. If the worker fears discharge, which could result from a too-low level of effort, extrinsic incentives are at work. Efficiency-wage theory fits: if an employer pays above-market wages, the threat of dismissal provides motivation. Similar remarks apply when promotion is possible and depends on a reading of the quality of one's work. Peer pressure can provide implicit and vague but still extrinsic incentives, if a slacker risks the opprobrium of his fellow workers.

Thus what is called intrinsic motivation may be (at least in part) the worker's response to fuzzy extrinsic motivators, such as fear of discharge, censure by fellow employees, or even the desire for coworkers' esteem (Bernheim, 1994). Because these motivators are fuzzy, observers may not see them and may misattribute their consequences to "intrinsic motivation."

Second, and more provocatively, the "disutility of effort" commonly found in simple agency models may be entirely wrong (James Baron, 1988). Workers may take sufficient pride in their work so that effort up to some level increases utility. How and why might this happen? Answers involve looking into the utility functions of individuals, terra incognito for standard microeconomics, so I leave this question hanging for now.

III. Economic Rationales that Depend on Preexisting (Vague) Incentives

Instead, I return to the stylized fact and look for rationales that play off the presence of preexisting albeit fuzzy incentives. Formal models of the rationales I offer will not be provided here, but I hope that formal models to illustrate these points will be obvious.

Jobs high in intrinsic motivation often involve a great deal of task ambiguity. Creativity is typically important, as is the quality of work. In short, the required tasks are multifaceted, with important facets that are hard to measure. In such situations, it can be tricky to get incentives right (Bengt Holmstrom and Paul Milgrom, 1991). An obvious rationale, then, is that the extrinsic incentives that are imposed—which almost necessarily will be relatively objective and formulaic—may be suboptimal, taking into account the full range of desired tasks.

Two interesting changes can be rung on this general theme. Individual workers often try to influence their superiors, spending valuable time in politicking or worse (Milgrom and John Roberts, 1988). When job tasks are ambiguous, forced evaluation/incentive formulas may invite efforts to corrupt the objective measures. Suppose a dean announces that she wants more innovation in the classroom; henceforth, departments will collect and use (in salary administration and in tenure reviews) statistics on percentage of new material in established courses, numbers of new courses offered, and so on. The predictable response is a lot of window-dressing "innovation," old courses given new titles and numbers, and many lunches for the department chair, where he is lobbied to count this or that bit of window dressing.

The second change to be rung on the multitask story invokes bounded rationality and unforeseen or uncontracted-for contingencies. When tasks are ambiguous and creativity is valuable, it is hard to say *ex ante* what should be done. Opportunistic responses to contingencies that arise are better than responses made to maximize some formula specified *ex ante.* Especially in smaller organizations, or where the evaluator is close to the work being evaluated, *ex post* evaluation can be constructed "fairly." Relying on evaluation criteria that are vague *ex ante* can thus give more powerful (better) incentives than criteria that are fixed formulaically *ex ante.*

Of course, a moral-hazard problem arises with any *ex ante* vague evaluation criteria. Generalized corruption is inevitably invited. Simple models suggest, however, that these problems can be mitigated, for example, where the principal/evaluator retains a significant stake in the economic health of the enterprise, where peer evaluation is used, and in small groups (where corruption is more likely to become known and dealt with by peer pressure).

So far the rationales suggested concern multitask jobs and possible misallocation of effort among tasks caused by extrinsic incentives. What about single-task jobs, with a single effort level? Suppose the worker wishes to keep his job. He will be evaluated by some criterion,

but is not sure what is the hurdle level for retention. Risk aversion can push him toward higher levels of effort, to be "safe"; when criteria are explicit and objective, he can put in just enough effort to stay safely employed. Moving to explicit criteria can lose the power of worker risk aversion, leading to a lower (average) effort level.

Formal models of this story, with an equilibrium model of the labor market, only half work. Workers subject to ambiguous evaluation criteria are worse off *ex ante* in consequence; retaining them (to match their reservation level of utility *ex ante* and to keep them from quitting for a better offer *ex post*) requires higher overall compensation. That is, we can get higher (average) levels of effort from workers by subjecting them to uncertain evaluation, but this does not improve the principal's bottom line.

Screening and signaling effects can be at work. Suppose some workers value autonomy more than others; others prefer strong economic incentives. If explicit extrinsic incentives are imposed on the work force, the mix of workers at the firm will of course change. If there is correlation between these tastes and specific worker abilities, a net drop in certain aspects of productivity could occur. Note that, if this rationale is correct, the effect of extrinsic incentives will work its way slowly and be associated with employee turnover.

To take a signaling story, most employers (even those who plan to cut and run to some foreign location) want their current employees to believe that a long-term, cordial employment relationship is in prospect. Thus an employer who truly plans to stay around may have to "oversignal" with incentive systems that are too expensive for those who plan to cut and run (i.e., incentive systems that are based on long-term monitoring, vague promotion criteria, etc.). Shifting to extrinsic (sharp) incentives may signal a change in plans to workers, who may respond with greater levels of opportunism, and the like.

Pleading space limitations, I will not pursue this class of rationales further. The point is, if "intrinsic motivation" is the response of workers to fuzzy but nonetheless extrinsic incentives, explicit extrinsic incentives that are imposed may fight rather than complement preexisting incentives. More generally, if adherence to norms depends on external enforcement [explanation (iii)], the impact of economic incentives on enforcement must be considered carefully.

IV. Changing the Individual's Utility Function

The stylized fact can also be rationalized by supposing that imposing extrinsic incentives changes the individual's disutility for the work involved; workers enjoy their work only in the absence of extrinsic incentives. Economists are loathe to rely on this sort of explanation, with good reason: it simply assumes the answer. Interesting applied theory can emerge by putting adherence to a norm into the utility function, if put there in an interesting fashion (e.g., as in Lindbeck et al. [1996], where the disutility from violating a norm rises with the level of adherence in the general population). But important questions are left unanswered.

To discipline the theory, one must dig deeper into how utility functions are determined. Excursions into cognitive and social psychology are warranted. To give an example, consider how social psychologists explain the stylized fact. Turning revealed preference on its head, the idea is that when a person performs some act, he looks for rationales that justify his actions. Specifically, if an employee undertakes some effort without the spur of some extrinsic incentive, he will rationalize his efforts as reflecting his enjoyment of the task. And since he enjoys it, he works harder at it. But if extrinsic incentives are put in place, he will attribute his efforts to those incentives, developing a distaste for the required effort.

The changing-tastes models of addiction seem in order, but with the addition that one does not become addicted if the drug is taken under external compulsion. As for normative lessons for economic incentives, an obvious one is that economic incentives, to complement intrinsic incentives, should emphasize the voluntary nature of the desired behavior. Recall, for example, that in Tracy Kidder's (1981) *Soul of a New Machine* what spurred on the troops was not pay-for-performance or the prospects of promotion, but "pinball effects"; if the group built a successful computer by working ridiculously long hours for

paltry pay, they would be allowed to do it again.

A second enter-the-utility-function rationale for the stylized fact is more social in nature, echoing themes sounded especially by Erving Goffman (1974) concerning role consistency. Accept that individuals are boundedly rational (or, at least, subject to costs of computation and cogitation) and that in specific relationships they thus try to fit the relationship to one of a few archetypes. In a kinship or family-like relationship, parties internalize each other's welfare, curbing their instincts to act opportunistically. In an arms-length, market relationship, caveat emptor is the rule.

Relationships within an organization, between employer and employee, or among employees, need not fit any particular archetype. But individuals, to make sense of them, will try to fit them into a standard pattern. Thus an employer who does not monitor closely the performance of employees (or, at least, does not appear to), and who complements this by symbolic acts of gift-giving, may engender kinship relations. This is not costless: forgiveness for misfeasance is probably higher than first-best. But when it is hard to provide strong economic incentives, where creativity or worker discretion is important, the benefits may outweigh the costs. Now imagine that this employer imposes a scheme of sharp extrinsic incentives. The nature of the relationship is muddied, at least, and the gain from the sharper direct incentives may be outweighed by the lost clarity in the relationship's nature. A worker who previously internalized the employer's welfare is sent signals that the relationship is a market exchange and reacts accordingly, taking fuller advantage of opportunities presented to him. Or, at least, he spends more time and effort trying to figure out what is appropriate in specific contingencies that arise (cf. Oliver Williamson's [1993] commentary on calculative and noncalculative trust).

V. Concluding Remarks

The interaction between norms and economic incentives will change, sometimes dramatically, depending on answers to the two italicized questions from the Introduction and others like them. Perhaps there are unambiguous answers. If so, they need to be discovered. I suspect, however, that answers will change with circumstances (e.g., internalization is more apt to be at work the longer one has adhered to a norm, as is internalization of the specific behavior). If this is right, there is a need to develop a better sense of what circumstances correlate with specific answers. Happily, because the different answers will give different theoretical interactions with economic incentives, once the theory is well developed there will be good grist for the empirical mill.

The results are likely to be messy. They will involve activities unfamiliar to economics (e.g., theories of how preferences are formed and reformed). But messy or not, they are important and must be pursued.

REFERENCES

Baron, James N. "The Employment Relation as a Social Relation." *Journal of the Japanese and International Economy*, December 1988, *2*(4), pp. 492–525.

Bernheim, Douglas B. "A Theory of Conformity." *Journal of Political Economy*, October 1994, *102*(5), pp. 841–77.

Deci, E. "The Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology*, April 1971, *18*(1), pp. 105–15.

Goffman, Erving. *Frame analysis*. New York: Harper & Row, 1974.

Holmstrom, Bengt and Milgrom, Paul. "Multi-Task Principal–Agent Analysis." *Journal of Law, Economics, and Organization*, Special Issue 1991, *7*, pp. 24–52.

Kidder, Tracy. *The soul of a new machine*. New York: Little, Brown, 1981.

Lazear, Edward P. "Performance, Pay, and Productivity." Mimeo, Stanford University, 1996.

Lepper, Mark; Greene, David and Nisbett, Richard. "Undermining Children's Intrinsic Interest with Extrinsic Reward." *Journal of Personality and Social Psychology*, October 1973, *28*(1), pp. 129–37.

Lindbeck, Assar; Nyberg, Sten and Weibull, Jorgen W. "Social Norms, the Welfare State, and Voting." Institute for International Economic Studies Seminar Paper No. 608, Stockholm University, 1996.

Milgrom, P. and Roberts, J. "An Economic Approach to Influence Activities in Organizations." *American Journal of Sociology*, Supplement 1988, *94*, pp. S154–79.

Staw, Barry M. "Intrinsic and Extrinsic Motivation," in H. Leavitt, L. Pondy, and D. Boje, eds., *Readings in managerial psychology*, 4th ed. Chicago: University of Chicago Press, 1989, pp. 36–71.

Williamson, O. "Calculativeness, Trust, and Economic Organization." *Journal of Law and Economics*, April 1993, *36*(1), part 2, pp. 453–86.