Removing financial incentives demotivates the brain

Colin F. Camerer[1]

Division of Humanities and Social Sciences and Computational and Neural Systems, California Institute of Technology, Pasadena, CA 91106

Social scientists and many biologists are all preoccupied in different ways with the nature and effects of the ways incentives influence behavior. One type of incentive is clearly intrinsic; it originates within a person and is often linked to exploratory behavior, hedonic pleasure from self-determined mastery, and desire to satisfy curiosity for its own sake (1). Another type of incentive is extrinsic; typically, it is designed and administered by an outside person or authority, is precise, and is usually financial, tied to fame, or has some other kind of monetizable status. Because of its various natures, intrinsic motivation is difficult to measure and observe. An adventurous step is to measure brain activity during conditions of apparent intrinsic and extrinsic motivation. Murayama et al. (2) do exactly this. Their results are striking evidence for a phenomenon often noted in social psychology—namely, extrinsic incentives (e.g., pay) can undermine intrinsic incentives (e.g., fun).

Neural Evidence of Incentive Undermining

To explore this undermining effect neurally, Murayama et al. (2) create a paradigm in which subjects must press a button 5 s after the end of a brief stopwatch (SW) cue. The task is engaging enough that subjects often do it for fun during a free-choice period after they are scanned.

The control subjects never received financial rewards for accuracy (and perhaps importantly, also did not know that other subjects did receive rewards). The treatment subjects received bonus rewards for accuracy in their first session, but no rewards in the next (removed incentives) session. The typical undermining effect is observed in behavior during free-choice periods after the scanning: the reward treatment subjects played the SW game less often.

A 2×2 ANOVA showed significantly greater blood-oxygen dependent level (BOLD) response to winning in the reward treatment compared with the control treatment in brain regions previously associated with reward. Specifically, BOLD signal responses to win trials (compared with lose trials) show a clear increase in BOLD level in the anterior striatum (caudate head) and the midbrain (which is known to project dopaminergic neurons to the striatum and prefrontal cortex). The results show that adding extrinsic incentives does increase reward-related response to wins (and subjects were more accurate in that condition). More interestingly, in the second session when the extrinsic incentive was removed, BOLD signal disappears in response to wins in the treatment condition. A similar pattern is shown in lateral prefrontal cortex (LPFC) when the task cue is first presented, indicating an SW trial (instead of an unengaging watch-stop control trial simply requiring a button press). The LPFC activity is sensibly interpreted as cognitive preparation associated with higher or lower motivation (3).

The piece de resistance is a final cross-subject neurometric analysis linking brain activity to postscanning behavior. Murayama et al. (2) find a significant correlation between the principal component of win-response BOLD reductions in three regions (because of removed incentives) and postscanner choice to play the SW game. This indicates that undermining is stronger for some people than others, and it might be possible to identify the strength of such an effect ex ante from functional magnetic resonance imaging (fMRI) (or another measure such as EEG, because LPFC is easy to localize in that way).

Increased incentives should be applied carefully, because removing them might damage or destroy a preexisting intrinsic incentive.

Understanding the neural basis (and robustness) of undermining is extremely important for both practical and scientific reasons.

The phenomenon is important in practice, because small concrete incentives are now being tried in a wide variety of domains such as public health and schooling. Most studies find that small concrete financial incentives, which reward good educational habits (prizes for reading books or good attendance), can change behavior substantially (4). Some of these positive results fade out over time, but some persist (contrary to the undermining prediction). One interpretation is that concrete incentives act like instructions, which both tell students exactly what attainable actions they should be doing and motivate them to act (5).

Scientifically, which incentives work to motivate people, and why, is a topic that cuts across all social sciences. Sociologists, anthropologists, political scientists, and economists emphasize the incentive effects of norms and roles, culture, elections and laws, and prices, respectively. These different emphases often lead to contentious debates and even to completely opposing recommendations.

Economists as a group espouse enormous faith in financial incentives. One popular book (6) boldly states, "Incentives matter. The literature of economics contains tens of thousands of empirical studies verifying this proposition, and not one that convincingly refutes it" (p. 9 in ref. 6). In the conduct of economics experiments, it is de rigeur to use financial incentives, because incentives offered by simply keeping score "are likely to be weak, erratic, and easily dominated by transactions costs, and subjects may be readily satiated with 'point' profits" (p. 277 in ref. 7). One recent imaging study does indicate that real consumer purchases activate reward regions more strongly and broadly than hypothetical choices (8), providing some support for experimental practice.

The undermining shown by Murayama et al. (2) is actually not inconsistent with the hypothesis that incentives matter; instead, it simply shows that removing an extrinsic incentive can do harm. One of the original experiments (9) on undermining showed that, when children colored pictures for fun, introducing a crisp extrinsic financial incentive—money per picture colored—had two interesting effects. First, the added extrinsic incentive would increase effort and productivity in the short run (sometimes at the expense of quality). This effect suggests that adding the extrinsic incentive on top of the intrinsic incentive had a positive effect. Second, when the extrinsic incentive was removed, effort and productivity then fell to a level below the original level. This change suggests that the extrinsic incentive actually permanently eroded the original

Author contributions: C.F.C. wrote the paper.

The author declares no conflict of interest.

See companion article on page 20911.

[1]E-mail: camerer@hss.caltech.edu.

intrinsic incentive or "crowded it out" (the phrase used by economists). The important lesson is that increased incentives should be applied carefully, because removing them might damage or destroy a preexisting intrinsic incentive.

Theories of Incentive Undermining

Although undermining is not new, it is also not well-understood (10). An early psychological explanation was that people perceive their motivation as caused solely by extrinsic incentives when they are introduced, and therefore, when those incentives are removed, their perceived motivation is removed also. A related view, called cognitive evaluation (11), is that the coercive extrinsic incentive somehow extinguishes self-determination. Until these accounts are fleshed out (actually, brained out) in terms of their associated neural computations, it is hard to say whether the neural evidence supports one view or another. It is possible that self-perception or reduced self-determination is happening upstream and influences signals in striatum, midbrain, and preparation in LPFC. If so, we should, in principle, be able to image a more general neural circuit including other activity in future studies.

Economists have also struggled to make sense of undermining using standard mathematical tools of rational choice. One explanation is that a lot of apparent intrinsic motivation is actually a response to fuzzy extrinsic incentives (which are difficult to observe and hence, are overattributed to intrinsic motivation). If so, then introducing clearer extrinsic incentives can remove the fuzzy ones and potentially reduce overall incentive (12). For example, a worker might be unsure what is expected of her to get ahead, and therefore, she works extra hard to be sure to reach that fuzzy goal. When the goal is made clear (e.g., a tangible promotion target), she can afford to work less hard, because the risk of falling short of a fuzzy goal is gone. Another theory is that an extrinsic incentive established by an authority conveys information about the job's difficulty or the worker's ability, which can reduce motivation (13). These economic explanations are meant to apply to workers in firms, and therefore, they do not offer a particularly persuasive account of a much simpler behavior, like the SW task; however, they do deserve much more attention nonetheless.

Another theory rooted in economics is that people do not know what the attainable reward (or wage) is for different tasks. When a reward level is established, it becomes a reference point for what is expected. When the reward is reduced, people withdraw effort in a microstrike to get the wage restored. A behavioral prediction of this theory is that, over time, the reward value and intrinsic motivation will return as people realize that the reward will not be restored, and they will stop striking. This account fits with the reduction in LPFC activity after incentives are removed but does not easily explain the fact that striatal and midbrain responses to wins are also lower when incentives are removed. Neurally, this theory predicts an increase in regions associated with perceptions of unfairness in bargaining when the incentive is removed [such neural activity is not examined by Murayama et al. (2) but is becoming understood in other studies (14, 15)].

The most promising explanation, noted by Murayama et al. (2), is that the brain adaptively codes reward based on recent reward levels and ranges, fictive outcomes, social comparison, and other yardsticks (16). When an extrinsic reward is removed, the residual intrinsic motivation seems smaller than if there were no extra reward in the first place. Thus, the experimental treatment group that underwent removed incentives would perceive less intrinsic motivation than control subjects who never received extra rewards at all. Besides massive evidence of adaptive coding in perception and psychophysics, there is now substantial evidence for this kind of adaptive coding in reward from behavior, fMRI (16), and single-unit recording (17, 18).

There is clearly plenty more to understand about the nature, persistence, and implications of the undermining effect. Murayama et al. (2) make large strides in this simple paper by showing both a reduction in apparent reward coding of wins (in striatum and midbrain) from removed incentives and a corresponding reduction of cognitive preparation in LPFC. These aggregated responses also reliably predict postscanner task motivation across subjects in a remarkable neurometric match of brain activity and plainly observable behavior. Their wonderfully simple time-guessing SW task works well to generate enough intrinsic incentive that undermining can reliably occur. However, it would be useful to explore a variety of other tasks and longer time periods (to see whether intrinsic incentive is ever spontaneously restored). Exploring the neural bases of the economists' theories about displacement of fuzzy extrinsic incentives and incentives signaling difficulty or skill would be worthwhile as well. Additionally, as their cross-subject neurometric correlation shows, there are individual differences in response to removed incentives, suggesting possible further influences of genes, age, culture, description framing, etc.

ACKNOWLEDGMENTS. Support for neuroeconomics in our group was provided by the Lipper Family Foundation, the Betty and Gordon Moore Foundation, and Tamagawa Global Center of Excellence grants.

1. Kang MJ, et al. (2009) The wick in the candle of learning: Curiosity activates reward regions and enhances memory for surprising facts. *Psych Sci* 20:963–973.
2. Murayama K, et al. (2010) Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proc Natl Acad Sci USA* 107:20911–20916.
3. Jimura K, Locke HS, Braver TS (2010) Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proc Natl Acad Sci USA* 107: 8871–8876.
4. Kremer M, Miguel E, Thornton R (2009) Incentives to learn. *Rev Econ Stat* 91:437–456.
5. Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J Risk Uncertainty* 19: 7–42.
6. Landsburg SE (1995) *The Armchair Economist: Economics and Everyday Life* (Free Press, New York).
7. Smith VL, et al. (1976) Experimental economics: Induced value theory. *Am Econ Rev* 66:274–279.
8. Kang MJ, et al. Hypothetical and real choice differentially active common value regions. *J Neurosci*, in press.
9. Lepper MR, Greene D, Nisbett RE (1973) Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. *J Pers Soc Psychol* 28:129–137.
10. Deci EL, Koestner R, Ryan RM (1999) A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psych Bull* 125: 627–668.
11. Ryan RM, Mims V, Koestner R (1983) Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *J Pers Soc Psychol* 45:736–750.
12. Kreps D (1997) Intrinsic motivation and extrinsic incentives. *Am Econ Rev* 87:359–364.
13. Benabou R, Tirole J (2003) Intrinsic and extrinsic motivation. *Rev Econ Stud* 70:489–520.
14. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300: 1755–1758.
15. Fehr E, Camerer CF (2007) Social neuroeconomics: The neural circuitry of social preferences. *TICS* 11: 419–427.
16. Seymour B, McClure SM (2008) Anchors, scales and the relative coding of value in the brain. *Curr Opin Neurobiol* 18:173–178.
17. Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307: 1642–1645.
18. Padoa-Schioppa C (2009) Range-adapting representation of economic value in the orbitofrontal cortex. *J Neurol* 29:14004–14014.