Three accounts of intrinsic motivation in economics: a pragmatic choice?

Blaž Remic

To cite this article: Blaž Remic (2021): Three accounts of intrinsic motivation in economics: a pragmatic choice?, Journal of Economic Methodology, DOI: 10.1080/1350178X.2021.1952291

To link to this article: https://doi.org/10.1080/1350178X.2021.1952291

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 13 Jul 2021.

Submit your article to this journal

Article views: 721

View related articles

View Crossmark data
Three accounts of intrinsic motivation in economics: a pragmatic choice?

Blaž Remic
Erasmus School of History, Culture & Communication, Rotterdam, Netherlands

ABSTRACT
This paper argues that the concept of intrinsic motivation has been used by economists in inconsistent ways because the underlying theories of intrinsic motivation, imported into economics from psychology, are competing and mutually exclusive despite employing the same terminology. I first identify and analyze three distinct economic accounts where intrinsic motivation refers to different things due to different underlying psychological theories employed. I then discuss implications these differences have for empirical work and incentive-based policy interventions. Finally, I use this discussion as a case study to demonstrate the shortcomings of the recently proposed pragmatic synthesis between neoclassical and behavioral economics. If there are multiple and fundamentally different psychological theories of the same phenomenon, using their insights in economic analysis is hardly just a matter of a straightforward pragmatic choice among the various tools in the economist’s toolbox.

1. Introduction
In recent decades, one of the cornerstone concepts in the literature at the intersection of economics and psychology has been intrinsic motivation. The concept of intrinsic motivation was developed in psychology as part of a critique of behaviorism and psychologists generally define it as the motivation for activities that are not means to some further end but an end in itself (e.g. Deci, 1971). Contrary to behaviorism’s straightforward positive relationship between reward and performance, the idea of intrinsic motivation is built around the hypothesis that when the activity is pursued as an end in itself rewards may have hidden costs in terms of being ineffective—or even deterrent—for the subsequent performance (Lepper & Greene, 1978). For economists, this hypothesis represents an especially intriguing puzzle since it seemingly violates the law of supply by implying that an increase in price may lead to less quantity supplied. The classic example is blood donation, where the claim is that introducing monetary compensation will result in less donations (Mellström & Johannesson, 2008; Titmuss, 1970). Some of the empirical applications of this hypothesis that helped to firmly establish it in the economic literature can be found in the analysis of tax compliance (Pommerehne & Frey, 1992), environmental morale and public virtue (Frey & Oberholzer-Gee, 1997), principal-agent relations (Bénabou & Tirole, 2003), prosocial behavior (Bénabou & Tirole, 2006), artistic production (Throsby, 2000), work in the care sector and public services (Nelson, 2006), and within the discussion on piece-rate pay vs. fixed wages (Gneezy & Rustichini, 2000). Apart from accounting for negative
effects of incentives, intrinsic motivation can also have a role as a sorting mechanism for workers, whereby employers can tease out intrinsically motivated workers by offering a wage below the competitive standard (Bohnet & Oberholzer-Gee, 2002; Heyes, 2005; Prendergast, 2008; for a criticism, see Nelson & Folbre, 2006).

While intrinsic motivation has been recently increasingly discussed in economics (notable examples are Besley & Ghatak, 2018; Bénabou et al., 2018; Bowles & Polanía-Reyes, 2012; Dellavigna & Pope, 2018; Festré & Garrouste, 2015; Le Grand, 2010; Luttmer & Singhal, 2014; Sugden, 2018), a proper systematic account of the theoretical underpinnings is still missing. Empirical work often proceeds based on ad hoc applications that only superficially delineate the underlying psychological mechanisms. Ashraf et al. (2014), for example, code work motivation by hair stylists as intrinsic if the subjects report they are doing it to ‘make people look nice’ and extrinsic if their reason is to ‘make money’ (4), which reflects the impression shared by many economists that intrinsic motivation refers to motivation that is negatively correlated to (or independent of) financial incentives. However, this is at odds with the above-mentioned general definition from psychology that intrinsic motivation is about activities that are not means to some further end but an end in itself. Both making people look nice and making money may be seen as separate pursuits with respect to the activity of cutting hair. This implies that in both cases the activity is not performed as an end in itself. Furthermore, the conceptual issues are sometimes entirely sidestepped. For example, DeVaro et al. (2017) do not provide a single definition of intrinsic motivation and do not cite any foundational research; they proceed with their empirical analysis based on a tacit assumption that the meaning and significance of the term intrinsic motivation is self-evident and uncontested.

The aim of this paper is twofold. In Section 2 I will first demonstrate that economists who imported the concept of intrinsic motivation from psychology into economics have been—often unknowingly—drawing on competing underlying psychological theories. In psychology, the counterintuitive hypothesis that extrinsic incentives might have detrimental effects was developed, tested, and published independently by three research groups in the early 1970s (Deci, 1971; Kruglanski et al., 1971; Lepper et al., 1973). This work initially culminated in a joint edited volume The Hidden Cost of Reward (Lepper & Greene, 1978) with contributions from all three groups. After that, however, their research programs diverged. The volume remains one of the key references for the detrimental effects of incentives, but what gets overlooked is that the three groups that contributed to it in fact offer distinct theories of intrinsic motivation. As a result there is not only a fundamental dispute with behaviorists about the correct theorizing and interpretation of the empirical findings concerning the relationship between rewards and subsequent performance; the research program of intrinsic motivation is divided internally as well. As Kruglanski et al. (2018) assert, ‘despite identical nomenclature [our models] refer to entirely different explananda’ (167, emphasis added). The economic literature mostly neglects this fact, and hence has imported a contested notion rather than a unified concept. Section 3 discusses what is at stake when considering empirical work and incentive-based policy interventions.

The second aim of the paper is to argue (in Section 4) that the discussion about the different accounts of intrinsic motivation has important consequences for the recent methodological debate concerning the relationship between neoclassical and behavioral approaches in economics. Most accounts of behavioral economics assume a relatively straightforward marriage of behavioral insights and rational choice theory. Sometimes this is even made explicit, and a ‘pragmatic’ stance is adopted suggesting that more realistic (i.e. psychologically plausible) behavioral assumptions can be applied (or added-on) whenever they prove necessary (Angner, 2019; Chetty, 2015). Given that within the pragmatic approach psychological plausibility explicitly corresponds to realistiness of the assumptions, this paper argues that a pragmatic stance towards choosing among different psychological insights when doing economics is often misguided because of the fundamentally different psychological underpinnings of the resulting integrated accounts. I demonstrate that such pragmatic application of intrinsic motivation insights to the economic analysis has resulted in a curious case of importing competing, contradictory and mutually exclusive theories of intrinsic
motivation from psychology into economics. It gave rise to multiple accounts of the same phenomenon. Arguably, the pragmatic attitude is in danger of ending up as a case of a particularly bad kind of ad hocness, leading to results that are closer to a set of descriptions than to real explanation.

2. Three accounts of intrinsic motivation in economics

Within economics we can distinguish three distinct explanatory accounts of the detrimental effects of incentives on intrinsic motivation: (i) the motivational account (Frey, 1997); (ii) the signaling account (Bénabou & Tirole, 2003, 2006); and (iii) the allocational account (Holmström & Milgrom, 1991). They differ with respect to the underlying causal mechanism, but the full extent of their systematic differences is poorly, if at all, recognized in the literature. For example, Bowles and Polanía-Reyes (2012) include (i) and (ii) in their comprehensive paper surveying the literature addressing the problem of substitutability of incentives and social preferences. But they nevertheless discuss the topic of intrinsic motivation exclusively in terms of the psychological mechanism underlying (i). Although Kreps (1997) and Festré and Garroutse (2015) include (iii) as an alternative account of the problem of the hidden costs of rewards, and Fehr et al. (2001) discuss it explicitly in terms of intrinsic motivation, this approach has been somewhat lost in the recent scholarly work as a distinct way of thinking about the problem of intrinsic motivation with its own psychological underpinnings. We will now look at (i) to (iii) in turn.

2.1. The motivational account: are your psychological needs met?

The motivational account is based on the motivation crowding theory (Frey, 1997; Frey & Jegen, 2001) that posits a possible downward sloping supply curve for effort by introducing the crowding out effect as an additional force working in the opposite direction than the relative price effect. In other words, an external intervention in the form of a reward may crowd out intrinsic motivation and thus result in less rather than more effort.

The concept of intrinsic motivation was for the first time explicitly introduced into economics by Stroebe and Frey (1982). They follow psychologists in characterizing intrinsic motivation as a motivation for performing an activity when ‘there is no reward except the activity itself’ (Stroebe & Frey, 1982, p. 121). In such a case, they argue, motivation is directed to the satisfaction of a psychological need. Crucially, as they point out, the self-interest postulate is not contradicted and thus the utility calculus essentially remains the same. What gets tweaked is the content of the personal utility function that in the motivational account acquires an additional expression in terms of the satisfaction of psychological needs, in particular the inner need for autonomy. In the motivational account, intrinsic motivation is modeled as a particular type of incentive for achieving a desired psychological state (Frey & Jegen, 2001, p. 590). And because the state of being intrinsically motivated is associated with higher productivity and willingness to exert effort (Grant, 2008), fluctuations in intrinsic motivation have a clear economic relevance.

The motivational account is often portrayed as a critique of the standard economic account based on the narrow notion of a rational economic agent. However, Frey’s earlier work reveals that he conceived it not as a critique but as an upgrade in the spirit of Becker’s economic imperialism (Romaniuc, 2017). In his initial foray into psychologically informed economics, Frey explicitly asserts that economic man and psychological man share the same utilitarian roots, both ‘assumed to respond systematically to positive and negative incentives’ (Stroebe & Frey, 1980, p. 120), and that both disciplines would largely agree that ‘man tries to behave rationally, i.e. chooses the action alternatives which are likely to be associated with the highest overall utility’ (Stroebe & Frey, 1980, p. 127). Stroebe and Frey (1980) point out that Becker’s economic approach to human behavior is a largely content-less theory and argue that psychology provides the needed empirical content that makes it possible to develop testable hypotheses that are based on more than monetary income alone.
To fully make sense of this model, we therefore need to understand the psychological theory underlying the motivational account: the self-determination theory (henceforth SDT; Ryan & Deci, 2017). The theory proposes that people have an inner need for autonomy and self-determination. This need is met when they perform an activity autonomously without a sense of being controlled from the outside: namely, when they are intrinsically, rather than extrinsically, motivated. SDT builds on the previously developed notion of the perceived locus of causality (DeCharms, 1968), according to which it matters whether the reward is perceived as external (that is, controlling), or internal (that is, when the person perceives herself as the origin of her behavior). The presence of a reward that is perceived as controlling will negatively affect the sense of autonomy and shift the locus of control to the outside, thus substituting extrinsic for intrinsic motivation. Such crowding out is not categorical, however, but happens on a scale representing a continuum between autonomously motivated and controlled behaviors (Ryan & Deci, 2017, p. 14).

Autonomy, referring to ‘the need to self-regulate one’s experiences and actions’ (Ryan & Deci, 2017, p. 10), is in SDT conceptualized as one of the three basic innate psychological needs, the other two being the need for competence (‘to feel effectance and mastery’; Ryan & Deci, 2017) and the need for relatedness (‘feeling socially connected’; Ryan & Deci, 2017). The motivation directed at fulfilling them is labeled intrinsic motivation. The crucial part of this theory—and also the core of its critique of behaviorism—is that fulfillment of the three needs cannot be elicited by external stimuli. Even more, any attempt to do so will be harmful because it will undermine intrinsic motivation and thus prevent the needs for autonomy, competence and relatedness to be met. Importantly, this is not to say that rewards have an immediate detrimental effect. As Ryan and Deci (2017) emphasize, ‘the scientific problem here is specifically their impact on the maintenance of intrinsically motivated behavior over time’ (127). The experimental evidence suggests that when a reward is first introduced and then taken away, the initially intrinsically motivated effort does not return to its before-treatment baseline. In other words, rewards lead to long-term corrupting effects. The shift of the perceived locus of control, brought about by extrinsic incentives, thus causes long-lasting damage to the psychological well-being of the individual.

It is important to stress that the three psychological needs of autonomy, competence and relatedness are in SDT understood as ‘nutrients that are essential for growth, integrity, and well-being’ and, as such, are ‘objective’ phenomena in that their deprivation or satisfaction has clear and measurable functional effects, effects that obtain regardless of one’s subjective goals or values’ (Ryan & Deci, 2017, p. 10, emphasis in original). These effects are akin to the effect that vitamin C deprivation will have on the body regardless of whether one believes in it or not. Furthermore, this functional relationship is an adaptive trait of the organism. As Ryan and Deci (2017) emphasize, the assumption of SDT is not that social-contextual events ‘cause’ intrinsic motivation—on the contrary, intrinsic motivation is understood as an evolved and inherent human propensity. The ultimate causes of intrinsic motivation lie in the selective advantages this propensity yielded in human prehistory. (124)

These advantages have to do with the inherent satisfaction brought about by an intrinsically motivated activity, which contributes to human learning and overall flourishing. Some economically relevant examples are experimentation and voluntary action (driven by the need for autonomy) that bring about innovative ideas; sustained play and persistence (driven by the need for competence) that have an important role in learning various skills and getting things done; and forming and sustaining communities (due to the need for relatedness) that is crucial for cooperation.

This discussion enables us now to provide a more detailed account of why we should expect voluntary blood donations to fall after an extrinsic reward has been introduced. The motivational account predicts that an extrinsic incentive will have measurable negative psychological effects on the individual person by depriving her of the ability to act autonomously, that is, out of intrinsic motivation. The psychological need for autonomy will not be met. But since, as we saw earlier, crowding out is not categorical, the extent of it will be determined by the strength of the extrinsic incentive. This implies that a large enough incentive will completely crowd out intrinsic motivation.
to donate blood out of an inner impulse to act altruistically. Once the intrinsic motivation to donate blood is crowded out, blood donation would be completely dependent on the extrinsic incentive. In the short term, this could even mean that an adequate total supply of blood for transfusions might be secured by offering a high enough monetary reward. However, the motivational account implies that such crowding out of intrinsic motivation will have important long-term costs in the form of negative effects on the psychological well-being of donors.

The motivational account thus puts at the center the question of personal psychological well-being, specifically the well-being that stems from the individual ability to act in a self-determined way. It is therefore not surprising that in terms of the explananda more specifically the motivational account prominently seeks to explain work morale and virtuous, moral, or prosocial behavior (Frey, 1997), which are the areas where the supply of activity by the agents is considered to be positively related to the ability to act out of inner impulses. The scope of the account is thus very much tied to the psychology of individual actors and the more general societal regularities are explained to the extent that they can be considered as aggregates of individual psychological reactions.

2.2. The signaling account: you are what you do

The signaling account models the interactive process in which actions are potentially perceived by others—or, in the case of self-signaling, by the agent himself—as displaying intrinsic motivation. It thus translates a motivational problem into a signaling problem. Intrinsic motivation is modeled as an incentive to achieve a desired reputational effect: being intrinsically motivated is a salient signal that people try to convey either by refusing to accept rewards or by diminishing the effort in the presence of rewards.

In the earlier incarnation of this account, Bénabou and Tirole (2003) developed a model of crowding out of intrinsic motivation based on the principal-agent framework with asymmetric information, where the principal has some private information about either the task or the ability of the agent to perform that task. The challenge for the agent is to strategize about the best course of action given the information about this private knowledge that he can elicit by reading the clues coming from the principal's actions. In such a situation, the offer of explicit incentives may signal possible 'bad news' to the agent: perhaps incentives are offered because the task is boring; or perhaps because the principal does not fully trust him. This will change the agent's information structure about his own abilities and, consequently, undermine his self-confidence. Crucially, the underlying mechanism of this model is based on an assumption that 'the agent will undertake the task only if he has sufficient confidence in his own ability to succeed, and in the project's net return' (Bénabou & Tirole, 2003, p. 491). Offering explicit incentives will thus result in reduced effort. However, recasting intrinsic motivation in terms of a rational calculation about the probability of success is the Achilles heel of this initial attempt, for it seemingly implies that what makes one intrinsically motivated is the increased chance of an extrinsic reward.

The signaling account was given its mature shape in Bénabou and Tirole (2006). They gave up on their initial attempt to model intrinsic motivation directly. Instead, they build on the assumption that people buy self-esteem by signaling some desired social trait, such as being intrinsically, rather than extrinsically, motivated to do something good for others. Crucially, the state of being intrinsically motivated does not play the decisive role in the deterring effects of incentives. The key assumption of the signaling account is the fundamental uncertainty with respect to the underlying motivations. This uncertainty applies to both the audience's ability to observe the true motivation of the agent, as well as to the agent's ability to directly know his or her own underlying motivations. Therefore, intrinsic motivation is not just an unobservable; it is also an unknown. As such, it always has to be inferred from actions. Thus, what matters in the model, as Bénabou and Tirole (2006) emphasize, is that the agents 'value being perceived, or perceiving themselves, as having high [intrinsic motivation]' (1657, emphasis added). This valuation—and not the motivational state as such—is what in the model determines the outcomes. Crowding out potentially happens due to reputational concerns...
regarding the inferences other people (or oneself, in the case of self-signaling) might draw about the supposed motivation for a particular act.

To capture that, Bénabou and Tirole (2006) propose a trichotomy of intrinsic, extrinsic, and reputational motivation. The latter relates to ‘what a person’s behavior says about him or her, which will depend on the informational and economic context, including what others are doing’ (Bénabou & Tirole, 2006, p. 1658). While the person in the model derives marginal benefit from all three factors in the mix, only the effort as the sum of the three motivations can be observed, not the particular mix. Therefore, the practical problem of strategic interaction is one of signal-extraction, where ‘rewards act like an increase in the noise-to-signal ratio, or even reverse the sign of the signal’ (Bénabou & Tirole, 2006, p. 1654). Incentives create noise in the inference about the underlying motivation because the audience will question whether the person did something for money rather than out of intrinsic motivation. This will, in turn, reduce the incentive to perform the activity in the presence of explicit incentives. As Bénabou and Tirole argue, ‘the presence of extrinsic incentives spoils the reputational value of good deeds, creating doubt about the extent to which they were performed for the incentives rather than for themselves’ Bénabou & Tirole, 2006, p. 1654. Thus, payments for blood donations do not have a direct corrupting psychological influence, as the motivational account would argue. What might lower the overall amount of blood donated is that payments remove the possibility for signaling virtue through blood donation. The person wants to be perceived as intrinsically motivated to donate blood, but in the presence of a reward she and others cannot be sure intrinsic motivation is indeed the right reason. In such a case, convincing others or herself that she is indeed intrinsically motivated requires her actions to clearly express the lack of a link between incentives and action, lest she taints the signal. As a result, in the presence of incentives she will not donate any blood at all. In other words, donations are reduced when people start to worry that getting rewarded will make them look as if they are doing it for the money.

The signaling account of intrinsic motivation in economics is a formalization of the overjustification hypothesis first put forward by psychologists Lepper et al. (1973). This approach explains intrinsic motivation in negative terms: in the absence of clear extrinsic rewards to account for given behavior, the person will attribute it to intrinsic motivation. Conversely, if rewards are clear and psychologically sufficient, behavior will be attributed to extrinsic motivation. The rather awkward name of the hypothesis refers to a situation where behavior initially attributed to intrinsic motivation will feel ‘overjustified’ once a reward is introduced. The presence of a reward will cause the person to reevaluate her engagement in the activity; and since now there is a clear link between activity and reward, the person will infer that the activity could not have been that intrinsically interesting after all. This will prompt her to reduce the effort when the reward is withdrawn again. Motivation is thus conceptualized as a kind of knowledge (Kruglanski, 1978, p. 24). For example, when children are given a reward for drawing—which is presumed to be intrinsically interesting—they will start to see the activity as a way to get the reward. This will lead to lower effort once the award is not presented, because children now simply have a different understanding of the reasons to do the activity. Important for our discussion is that the attributional mechanism does not depend on any particular nature of intervention: it is not about money as such. As Lepper et al. (1973) argue, ‘an overjustification effect is predicted for any situation which results in an extrinsic attribution where previously intrinsic interest was the only salient attribution’ (130).

Theoretically, this psychological approach is a combination of self-perception theory (Bem, 1972) and attribution theory (Kelley, 1967). According to the former, ‘individuals come to “know” their own attitudes, emotions, and other internal states partially by inferring them from observations of their own overt behavior and/or the circumstances in which this behavior occurs’ (Bem, 1972, p. 2). In other words, we cannot know ourselves directly, but only through indirect inferences. Attribution theory is a cognitive theory concerned with the processes of inference in general (Kruglanski, 1978, pp. 24–25), but with regard to intrinsic motivation it argues that people explain behavior— their own or others’—by attributing it to either internal or external reasons. Internal attribution means that one thinks of oneself as being the kind of person that behaves in the observed way;
and external attribution means that the causes of behavior are attributed to some situational factors. To be intrinsically motivated to donate blood then means that one perceives oneself to simply be the kind of person that donates blood without having to be paid for it. However, self-perception theory says that that can never be directly known. In the presence of an extrinsic reward, it will not be possible to clearly infer that one indeed is such a person, for even in the case of strong self-signaling it is impossible to be sure—perhaps the person is simply repressing the fact that she is doing it just for the money. Therefore, also in the case of self-signaling it is only reputation that is observable. In this sense, self-signaling is actually a model of self-deception (Mijović-Prelec & Prelec, 2010) and not a model of the underlying motivation. The combination of self-perception and attribution theories reveals the dynamic between, on the one hand, our inability to know ourselves and, on the other hand, the ways in which we nevertheless get to define ourselves by the actions we perform. In other words, actions serve as signals about the possible reasons and motivations for behavior. It is now obvious that this theory remains closer to behaviorism than the motivational account.

These signals are important for social interaction, since they enable the agents—individuals, but potentially also organizations—to convey and infer information that guide further action and allow coordination. Rather than what happens at the individual psychological level, of interest here are the consequences that reading other people’s actions and responding to them has for social interactions. The scope of the signaling account is thus rather indistinct, extending from the micro perspective of the principal-agent interactions to the larger-scale institutionalized rules that structure the incentives and serve as sources of information for strategizing about the possible actions of others and how certain actions will be perceived in the context of interaction. Illustrative examples thus encompass any situation where it is desirable to be perceived as doing something out of intrinsic motivation, such as contribution to public goods, donating to non-profit or charity, being altruistic, reciprocating a friendly action, refraining from imposing negative externalities on others, and more (see Bénabou & Tirole, 2006). Importantly, the scope is determined by the socio-cultural meaning and desirability of these acts, and not by inner psychological drives.

2.3. The allocational account: what are you really rewarded for?

Contrary to the first two, the allocational account developed by Holmström and Milgrom (1991) is not explicitly about the concept of intrinsic motivation. However, the fact that it does effectively represent a distinctive account of it has not gone unnoticed among economists (see, e.g. Fehr et al., 2001). The account focuses on the actor’s allocation of effort among the various aspects of the task to be performed. In the model, these aspects are called dimensions. The key factor is that the different dimensions are, for various reasons, differently incentivized. Intrinsic motivation can thus be modeled as a motivation for activity that is either not explicitly incentivized or is incentivized less, relative to other dimensions of the same task.

For example, the job of a university professor consists of dimensions such as research, teaching, administrative tasks, and networking with stakeholders outside of academia. On top of that, those dimensions can be further divided into subdimensions. Research activity consists of publications, generating new ideas, and serving the scientific community by doing peer review. However, it may be that it is harder or more costly to measure and evaluate the performance on some of these dimensions. In such a case it is likely that incentives will be tied to the dimension where performance standards can be more easily specified and monitored. However, when the employer starts to explicitly incentivize a particular dimension, employees will start to focus on it at the expense of the others. The provision of explicit incentives for a measurable dimension will raise the opportunity costs of exerting the effort on the dimension where performance is hard to measure. By engaging in the latter, the employee is foregoing the additional money he could be earning by focusing solely on the former. He will then reallocate the effort to the measurable dimension in order to maximize the financial benefits. Consequently, the motivation to exert effort on the non-measurable dimension will appear to be crowded out by some extrinsic incentive. If professors are evaluated and rewarded
primarily based on their publication record or their administrative effort, this may come at the expense of their engagement in teaching.

Such reasoning, however, may still seem like the standard microeconomic logic regarding the substitutability of inputs. What nonetheless makes this model specifically about intrinsic motivation is the assumption—made explicit by Holmström and Milgrom (1991) in a footnote only—that agents ‘are motivated to […] supply [some] inputs even without incentive pay’ (32n9). In other words, the absence of explicit incentives for a particular dimension does not mean that the agent will not exert herself. However, that effort can get completely reallocated to another dimension if that dimension is clearly incentivized. Holmström and Milgrom reason that it may, therefore, be better to reduce explicit incentives for measurable dimensions if the desired outcome is some effort on the unmeasurable component—which is analogous to reasoning that it is better not to introduce extrinsic incentives lest intrinsic motivation would get crowded out. However, the point is not that incentives as such are harmful. According to the allocational account, some form of incentive contract is always needed, since Holmström and Milgrom assume that the unincentivized supply will not reach the optimum level. But rather than performance pay, the optimal incentive contract in such cases would pay a fixed wage, because the latter does not provide explicit incentives for any particular dimension and thus avoids crowding out.

As with the other two accounts, it helps at this point to examine the underlying psychological theory implied by the allocational account: the theoretical take on intrinsic motivation as developed within the tradition of behaviorism in psychology.7 It must be said from the onset that Holmström and Milgrom do not build directly on this theory qua the psychological theory underlying their account. Nevertheless, its main points are tacitly accepted.

For our purposes, the most important aspect of the behavioristic approach to intrinsic motivation is that it does not see intrinsically motivated behavior as functionally different from extrinsically motivated behavior. They both follow the same logic of responding to some set of incentives to achieve a desired goal. However, the intrinsic/extrinsic dichotomy is not understood in terms of incentives being internal or external to the person, but as internal or external to the activity (Dickinson, 1989). We can illustrate this distinction by considering a puzzle from Rheinberg and Engeser (2018): ‘Why [do] some top-earning football and tennis players give up the game altogether when they retire, whereas some former professional skiers and world cup surfers continue to practice their sports enthusiastically, even without the prospect of material rewards’ (591)? Intrinsic motivation is here associated with the incentives related to the activity itself. Aside from the factors external to the activity, such as competition, fame, or money, the activity of surfing as such seems to be more engaging than playing football is. This relates to the part of the allocational account that assumes people will to some extent engage in the activity even without incentive pay. However, as pointed out earlier, such intrinsic aspects of the activity are—on their own—not enough to provide the optimal amount of effort. After all, it is not surprising that top surfers train the hardest and develop the most when the activity is part of a competitive context with clear financial rewards, and not after they retire from the competitive career.

For proponents of the behavioristic approach, the two motivations being functionally the same means that potential crowding out effects ‘lose much of their philosophical importance,’ yet ‘they remain empirically interesting’ (Dickinson, 1989, p. 12). This empirical interest lies in the challenge the empirical findings present to researchers: to keep on looking for the hidden incentive effects that are actually at work in a given case, rather than simply focusing on seemingly obvious rewards. As Cameron and Pierce (1994) state, ‘intrinsically motivated behavior is simply behavior for which appropriate controlling stimuli have yet to be specified’ (364). In the language of behaviorism: the reward that is offered may not be the actual reinforcer affecting the behavior. To recall the difference between rewards and reinforcers: a reinforcer is something that positively affects the behavior; a reward, on the other hand, is only assumed to have such a function, without the direct link being really shown (Cameron & Pierce, 1994, p. 364). Rewarding the teachers based on students’ results on the standardized tests may be meant to incentivize their teaching efforts, but
in reality, the real reinforcer at work is the mechanism that rewards teaching to the test. Therefore, it may be misguided to focus only on links between rewards and behavior, without a proper understanding of the nature of the task: the interplay between its many dimensions and incentives (perhaps yet unknown) that are present in the situation. This pertains all the more to complex and multifaceted forms of behavior—which behaviors high on intrinsic motivation are usually assumed to be—since behaviorists argue that in such cases it becomes increasingly difficult to identify the sources of control (Dickinson, 1989).

The allocational dynamic is thus not explained by referring to competing types of motivation, but with respect to competing reinforcers. One clear implication is that the allocational account does not see activities such as blood donation as being primarily intrinsically motivated in the first place. If so, it would probably mean that one is motivated by the activity of sticking a needle into one’s vein and experiencing pumping out the blood. For the allocational account, everything that is outside of such activity is extrinsic by definition. Altruism, for example, is clearly an extrinsic aspect of an activity, which is at odds with the fact that altruistic acts are often being portrayed as intrinsically motivated. Once we recognize this, we see why the most important question is not about the intrinsic or extrinsic motivation, but about the relative effects of competing incentives that lead to different allocations of effort.

This question, however, crucially depends on how the actor perceives the nature of the activity, since perception will define the relevant dimensions of the task. Consider that the actor in our exemplary case of blood donation perceives the task as consisting of the following two dimensions: (a) to incur costs in terms of pain and foregone work; and (b) to benefit unspecified others. Those two dimensions can be pursued simultaneously. But since the reward is meant to provide compensation for a costly activity, and because the benefit to others is much harder to measure than the individual cost, the reward will reallocate the attention from (b) to (a), because the latter is what is effectively incentivized. The fact that most countries have a compensation scheme in place for blood donors suggests that this mechanism is rather plausible. However, it also means that the dimension to benefit others will receive proportionally less attention as the reward gets larger.

3. The differences between accounts matter for empirical work and for incentive-based interventions

The preceding elaboration now enables us to more carefully examine what the differences between the approaches imply for economic research and for the design of incentive-based interventions. Perhaps the most obvious observation is that the three accounts employ different notions of intrinsicness. In line with the underlying psychological theories, intrinsic motivation is conceptualized either in terms of motivation for rewards internal to the person (motivational account); the absence of extrinsic rewards (signaling account); or rewards internal to the task (allocational account). This is not only of theoretical significance, but has especially important implications for empirical research, where appropriate and successful operationalization of intrinsic motivation depends on which account is adopted. When asking respondents to reflect on the reasons for their actions it matters a great deal what notion of intrinsicness researchers employ when coding and sorting the responses. Similarly, when explaining experimental results, the domain of behavior for which intrinsic motivation might be plausibly supposed is an important factor.

The second observation is that the three accounts imply different visions for the nature of the economic problem that they are addressing. It matters whether the economic problem is how to balance the incentives in order to achieve the optimal level of well-being (motivational account), to achieve the optimal level of social reputation (signaling account), or to achieve an optimal level of a multi-dimensional task performance (allocational account). Organizations or governments seeking to avoid crowding out through appropriate rewards or incentive schemes need to know which theory is the most appropriate, in order to properly function.
The feature of the motivational account that has probably proved to be the most attractive for practical applications is that it builds on a psychological account that theorizes psychological well-being. As such, it has an advantage of incorporating a wider notion of welfare than only pecuniary payoff. As pointed out above, it is a version of Gary Becker’s economic approach to human behavior. Becker refrained from giving his notion of stable preferences any objective form (see Emmett, 2006), but Frey’s account does precisely that. The utility brought about by the state of being intrinsically motivated is conceptualized in terms of the satisfaction of objective psychological needs. This is especially attractive for the study and interventions in the workplace, since it helps to disentangle the monetary compensation of workers and the actual feelings people have while working. It upsets the traditional notion that sufficiently high monetary compensation can buy any kind of hardships at work. The reason is that while compensation may indeed have short term benefits, it will nevertheless lead to long-term negative psychological effects by crowding out the intrinsic motivation. The needs-based motivational account implies an emphasis on the research into well-being. Not surprisingly, following this direction has led Frey to recently engage with the economics of happiness as a natural extension of this account (see Frey, 2018).

The focus in the signaling account on strategic interaction has great merit and value for analysis of any situation where bargaining or concerns over reputation play a decisive role. While the motivational account has trouble accounting for the adaptations people make in response to the changes in the social desirability of particular activities, the signaling account in this case provides more precision when describing the exact mechanism of the change. It takes ‘being perceived as intrinsically motivated’ as a desired outcome of the signaling process and then shows the dynamics of the decision making and adaptation when incentives interfere with the signals. It is not that people seek to signal something that is objectively related to their own individual psychological well-being; they seek to signal whatever is socially desirable, which adds a distinctly social component to the individual preferences. The agent is self-centered, but the preferences that enter his strategy calculus are completely socially constructed and entirely conditional on the preferences of others. Because the model is not directly about intrinsic motivation, the signaling account implies further research into social norms and values that drive what is considered to be a desirable trait, and how the relative social status affects those values. For policymakers or managers seeking to promote more intrinsically motivated behavior, simply raising the status of that activity might be sufficient.

The allocational account operationalizes intrinsic motivation in terms of the changing effort along different dimensions of the task. Thus, it forces us to think about the possible dimensions that we might be overlooking and to think about how incentives affect the performance along those dimensions. Contrary to the other two, the model does not directly require an additional type of motivation. This seemingly makes it immune to some of the issues that come along when extra psychological or reputational factors are added to the analysis. However, according to the allocational account, incentives only affect the marginal utility of effort, not the fundamental decision to engage in the activity. The reasons why people do what they do are not directly linked to the incentive schemes. There seems to be a fundamental element of purpose to the activity that does not depend on economic considerations. This implies, perhaps somewhat surprisingly, that the allocational account offers a fundamental modification in which work is no longer simply a cost. However, while the motivational account explicitly assumes that without the activity the person would be unhappy; and the signaling account assumes that acquiring social prestige in the form of being perceived as intrinsically motivated is central to living in the social world; the underlying assumption of the importance of the unincentivized supply of activity remains unexamined by the allocational account. While the latter is most closely aligned with the standard notion of utility in economics, this omission leaves a gap when it comes to explaining behavior in the first place. This paper provides some clue by linking allocational account with the behavioristic account of intrinsic motivation. Our discussion suggests that in order to better understand the
workings of incentives, the allocational account implies further research into how the actor perceives and interprets the nature of the activity.

The third observation that follows from our discussion is that contrary to what many economists think, the notion of intrinsic motivation does not simply refer to motivation negatively correlated to (or independent of) financial incentives. While much experimental work is indeed performed by using money as a proxy for extrinsic motivation, a more careful reading of psychological literature reveals that it is not about money as such. The motivational account is about any reward that feels controlling; the signaling account is about any reward that negatively affects the reputation; and the allocational account is about any reward that targets a particular measurable dimension of the task. Empirical research thus faces a much more complex challenge of identifying and tracing these incentive effects, be it that they are dependent on the individual psychology of the actor, on the social environment, or on the actor’s interpretive scheme about the task.

Last but not least, our fourth observation is that all three accounts point to a clear economic relevance of the crowding out effects. This leads to—perhaps somewhat ironically—a pronounced economic motive by policymakers and employers to intervene. The desire to formulate advice for interventions probably explains a large portion of the interest by economists in intrinsic motivation. By clarifying the conceptual confusion, this paper provides a more solid ground for developing policy proposals and workplace modifications, since our discussion shows that differences in the accounts imply fundamentally different nature of the proposed interventions. However, since the three accounts are built on different competing psychological theories of the same concept, the question that probably matters a great deal to policy makers and designers of incentives is: which theory is the right one?

4. Methodological implications: is it really a matter of pragmatic choice?

So far, the aim of our discussion has been to provide clarification and better understanding of the competing accounts of intrinsic motivation in economics. In this section, I want to demonstrate that the fact that there exist different theoretical accounts of the same concept has deeper consequences for the recent methodological debate concerning the relationship between neoclassical and behavioral approaches in economics.

Angner (2019) argues that a new synthesis has taken place between neoclassical and behavioral economics. His claim is built on an observation that economists are increasingly adopting what Chetty (2015) described as a ‘more pragmatic, policy-oriented perspective’ whereby behavioral factors are to be incorporated in the analysis ‘to the extent that they improve empirical predictions and policy decisions’ (1). However, contrary to the view held by many economists that such a synthesis of behavioral and neoclassical economics makes behavioral economics obsolete as a distinct research pursuit, Angner argues that the nature of the new synthesis is behavioral and not neoclassical. Economists, such as Chetty, who advocate for the pragmatic approach, are only able to legitimately do so by assuming a methodological position specific to behavioral economics that ‘it is legitimate and sometimes necessary to incorporate behavioral factors in economic models’ (Angner, 2019, p. 200). The pragmatic inclusion of behavioral factors is meant ‘to increase the explanatory and predictive power of economic theory on the margin’ (Angner, 2019, p. 199) by providing a set of realistic (or psychologically plausible) assumptions, where realistic is to be understood as ‘consistent with the best available psychology’ (Angner & Loewenstein, 2012, p. 642).

The choice to use the word synthesis is of course not arbitrary. It is an explicit allusion to the neoclassical synthesis, as well as to the neo-Darwinian synthesis, which have both been considered to be examples of explanatory unification (Kitcher, 1981; Mäki, 2001). Therefore, we can assume that the new synthesis has similar unificatory aspirations. It might perhaps be obvious that the new synthesis is meant to unify neoclassical and behavioral economics in order to increase explanatory power. However, Angner’s insistence on pragmatic attitude towards the use of assumptions raises important questions with regard to the nature of the proposed unification. To use the terminology
developed by Mäki (2001), unification can assume two forms: derivational and ontological. In the case of the former, what matters is that more is explained by using a smaller number of assumptions. In general, the tendency of neoclassical economics to favor (sometimes in silly ways) theoretical consistency over many other things reflects a high degree of derivational unification. Ontological unification, on the other hand, refers to the discovery that phenomena to be explained share the same underlying principles. Systematic application of psychological insights to derive behavioral microfoundations for macro theories would be an example of such an attempt. The new synthesis is clearly not a case of derivational unification, since the pragmatic attitude towards the use of behavioral and neoclassical assumptions as tools suggests that the goal of unification is not to decrease the number of axioms but rather to expand the available toolbox. Some problems do not need behavioral factors and they can be accounted for with the use of traditional tools; others will be explained in a better way by incorporating psychology. For similar reasons, the new synthesis also cannot be categorized as an example of ontological unification, since behavioral factors are not assumed to play a role everywhere. Pragmatism makes it hard to assess the nature and degree of unification implied by the new synthesis.

It may be argued, however, that this is precisely what the new synthesis is about; that it is so pragmatic as to only care about the internal consistency of the use of a concept in a given paper. However, offering such a large degree of freedom of choice to the practicing scientist when explaining the social world raises the problem of ad hocness of explanations. Indeed, one desirable property of a theory of explanation is that ‘it should be objective—what counts as an explanation should not depend on the idiosyncrasies and changing tastes of scientists and historical periods’ (Friedman, 1974, p. 14). Furthermore, as the classic example of Ptolemaic epicycles teaches us, increased predictive power is not enough to accept an ad hoc solution (Forster & Sober, 1994). Granted, the new synthesis is not pragmatic to this extent, and it surely does not imply that ‘anything goes.’ After all, behavioral factors are supposed to be realistic. However, the discussion in this paper shows that one major problem with this position is that the question of what is considered realistic—what for the pragmatic practitioner counts as the best available psychology—is not as straightforward as it may seem. Psychological theories are competing not only in terms of which one best explains behavior. They are also competing in terms of their fundamental view of human psychology. Ranking may thus be severely complicated. Do we have direct access to our motivational state, as SDT suggests? Or do we always have to infer it from our actions, as is suggested by the self-perception theory? It is hard to square these issues with the prescription to employ the best available psychological theory, because the choice here automatically entails a choice between two or more rather different underlying visions of human psychology. The pragmatic approach has a hard time dealing with competing psychological approaches of the same concept. The discussion in this paper should remind us that by ignoring the underlying differences there is a danger of sliding into ad hocness of the worst kind: only behavioral description, rather than explanation. Indeed, the new behavioral synthesis can increase explanatory power only if we assume that its psychological part is unified.

Thus, Angner’s view that the new synthesis is behavioral in nature suffers from the same problem that he identifies with respect to Chetty’s take on the pragmatic approach: they both assume certain answers to fundamental questions about the foundations of behavioral economics. Chetty assumes the answers that are the same as the one’s provided by behavioral economists. And Angner assumes that the clause about the best available psychology secures coherent and uniformly plausible psychological foundations. This highlights the problem of thinking about behavioral economics simply as an instance of importing psychological insights into economics to provide it with more realistic (i.e. psychologically plausible) assumptions. A synthesis that is pragmatic to the extent that it encompasses any model that combines economics with any psychological account can hardly be thought as explanatory. It is closer to being a set of descriptive accounts. More emphasis should be given to the fact that psychology is a plural discipline and that we thus have to be careful when commenting on seemingly similar psychological theories that have been imported into economics as part of a coherent behavioral economics approach.
5. Conclusion

This paper has argued that the concept of intrinsic motivation—imported into economics in attempts to explain why incentives may sometimes backfire—has been used in economics in inconsistent ways. We identified three distinct accounts of intrinsic motivation in economics and demonstrated how they differ in terms of the underlying psychological theories they employ, and in the explanatory mechanisms they provide for the potential detrimental effects of incentives. Understanding these differences helps clearing out the conceptual confusion that surrounds the concept of intrinsic motivation used in economics.

Economists are increasingly adopting a view that integration of psychological insights into economics is a pragmatic issue, guided by the nature of the puzzle that empirical data present to us. According to this view, models based on psychological insights are simply another tool in the economist’s toolbox. However, it gets complicated when we consider that the underlying psychological theories might not only be at odds with economics but competing within psychology as well. It is not only that we may wonder how to choose the appropriate tool. If a certain psychological model is adopted, the problem arises that this particular combination of economics and psychology may be fundamentally incompatible with combinations that rely on other psychological models. Arguably, the new behavioral synthesis cannot be based on fundamentally different psychological theories.

Intrinsic motivation provides a good exemplary case for demonstrating what is at stake, because the psychological underpinnings employed by the economic accounts may be either fundamentally at odds with the subjectivist economic approach (as in the case of the motivational account, which is based on an objective notion of human needs), they may lead to a reconceptualization of the nature of the individual in economics (as is the case in the signaling account that turns the individual into a seeker of social prestige and self-esteem), or they may enable economists to bypass any psychological issues (such as in the allocational account, which in terms of psychology remains an empty shell). In light of this, the pragmatic choice may lose much of its straightforwardness.

Notes

1. Rather than reward, behaviorists employ a technical term positive reinforcement, which refers to any stimulus that increases the probability of a response. Conversely, negative reinforcement—akin to punishment—refers to a stimulus that decreases the probability of a response (Skinner, 1953; for a discussion see McCullers, 1978, pp. 12–16).

2. Two of these three psychological accounts have been employed by economists and will be discussed later. The third one stems from the initial work by Kruglanski et al. (1971) and has recently been restated as a means-ends fusion model of intrinsic motivation (Kruglanski et al., 2018; see also Harackiewicz & Sansone, 2000). While this approach has recently received some attention in the management literature (e.g., Foss & Lindenberg, 2011; Woolley & Fishbach, 2018) there has not yet been any systematic attempt to adopt it within economics. Therefore, it will not be further considered in this paper.

3. They justify their decision to do so by arguing that “it is likely that [in (iii)] more than one mechanism is at work” (Bowles & Polania-Reyes, 2012, p. 398). Furthermore, they conceptualize (ii) in strictly informational and strategic terms without recognizing that the underlying logic of this account stems from a distinct psychological theory of intrinsic motivation that is different from the one they mention in relation to (i).

4. This reflects the fact that SDT consists of a set of “mini-theories,” one of those being cognitive evaluation theory, which focuses on the issues of autonomy and control (see Ryan & Deci, 2017). Cognitive evaluation theory was the first one to be formulated within this research program (e.g., Deci, 1975), and is what Frey primarily based his account on. While autonomy alone would arguably be enough to present the core of the motivational account, it is instructive to include a discussion of the other two needs as well, since the recent work within this account has been drawing on all three of them. Competence has been emphasized by economists discussing craftsmanship and artistic creativity (e.g., Klamer, 2016). Frey himself has been incorporating the need for relatedness in his more recent work on happiness (e.g., Frey & Stutzer, 2002; Frey & Gallus, 2012), but it arguably also applies well to the situations governed by social norms. Empirical work in economics based on SDT sometimes incorporates all three needs (e.g., Cnossen et al., 2019; Nikolova & Cnossen, 2020).
5. In a classic demonstration of this effect participants in two groups are assigned a task of solving puzzles (Deci, 1971). The experiment consists of three rounds, the only difference being that the treatment group receives a performance-contingent reward in the second round (but is told that there is no more money available for the third round). The results show that participants in the treatment group demonstrate lower levels of motivation in the third round compared to the control group.

6. Gold (2019, p. 173) provides a similar observation with respect to the inability of signaling approaches to explain the claims about the corrupting effects of commodification on the values people hold with respect to activities such as blood donation: signaling explanations do not model the underlying valuations but only the process of self-deception.

7. The psychological literature refers to this approach as behavioral approach. In order to avoid a possible confusion with behavioral economics I adopt the term behavioristic instead.

Acknowledgements

The paper has benefited from the discussions and feedback at the ESHET summer school in Lake Como, Econ and Culture seminar at the Erasmus University Rotterdam, and INEM conference in Helsinki. I am grateful to Erwin Dekker and two anonymous reviewers for their outstanding work in providing invaluable feedback and comments that helped to shape the paper. All remaining errors are mine.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Blaž Remic is a PhD candidate at the Erasmus University Rotterdam. His fields of interest include methodological implications of behavioral economics, philosophy of economics, and economic sociology. He is currently particularly interested in the relationship between rationality and motivation.

References


