Boris Begović, PhD*

Samuel Bowles, *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*, Yale University Press, New Haven & London, 2016, 272

In the very setting of the scene for the book, Bowles sends a clear signal: his aim is to convince a reader that *Homo economicus*, a model of amoral selfishness, is not a prudent assumption on which laws and public policies should be based. It is the incumbent legislative design based on that very assumption that creates incentives to individuals, to direct them towards some desirable behaviour. That is wrong, according to the author, at least in some cases, like examples, frequently and ubiquitously referred to in the book, of parents and their children in Haifa kindergarten and Boston firefighters and their Commissioner. In both cases introduction of (monetary) fines to suppress observed behaviour (late collecting of children and frequent sick leave) produced the opposite results: the intensity of the undesirable behaviour increased. That was apparently sufficient for a bold initial statement by the author: "Economists, who have placed the act of choosing at the center of all human activities, have now discovered, in short, that people are not very good choosers" (p. 8). And all that enlightenment thanks to Thaler, Sunstein, Kahneman, *et al.*, i.e. behavioural economics.

The bottom line, according to Bowles, is that modern legislation is based on the assumption that people are bad, traced back to Hobs and Machiavelli. In the words David Hume, quoted on the front page of the book "every man ought to be supposed to be a *knave* and to have no other end, in his actions, than his private interest". This line of reasoning, amply supported by Oliver Wendell Holmes, means that good legislation is the one made for bad people. Bowles disagrees – he proposes an analytical framework in which the conclusion is that this is not inevitable. The framework is simple: individuals behave according to (external) incentives and (internal) social preferences, i.e. moral norms. The former creates instrumental motives, the latter – intrinsic motives. If the two are

* Professor at the University of Belgrade, Faculty of Law, *begovic@ius.bg.ac.rs*.

separable, then good legislation should unconditionally be designed for bad people. Their desirable behaviour would be driven by the law and the incentives that it creates, and desirable behaviour of the good people will be driven by their moral norms. Bad people would not murder anyone because of the expected punishment, i.e. deterrent that it creates, while good people would not kill anyone because of the categorical imperative not to kill. In economics jargon, there is additivity of motives, since they are separable: one works independently of the other, but both work in the same direction.

The analytical problem emerges if they are not separable. In the case of superadditivity, or crowding in (synergy), incentives reinforce moral norms and the joint effect of the two of them is more than just adding the effect of one to the other – obviously not a policy problem. Labour contracts are incomplete, the incentives they create are not perfect for reaching Pareto optimality, but they reinforce work ethics and the outcome is better than just separate effects of incentives and moral norms. According to the author "Moral economy is not an oxymoron" (p. 35). The policy problem, nonetheless, exist in the case of subadditivity, i.e. crowding out (negative synergy) – the case when incentives undermine moral norms. Hence, incentives have countervailing effects in such a situation: the indirect effects of incentives undermine their direct effect. Bowles introduces the notion of strong crowding out, a situation in which the outcome is worse than the situation when incentives are not introduced at all. This is the bottom line of the Haifa case:[1] when a fine was introduced for parents who are late collecting their children, the parents started to collect their children even later than before!

Based on these considerations, a clear analytical framework is introduced with specific best response functions for each case and it is enhanced with the introduction of two types of crowding out. The first one is categorical crowding out, when the sheer presence of the incentives affects a person's social preferences. In other words, in the case of categorical crowding out, the magnitude of the incentive is not relevant. It is the introduction of the fine or subsidy itself that matters for crowding out, rather than their magnitude. The other one is marginal crowding out, in which case the magnitude of the incentive, i.e. its marginal value, is relevant for the outcome;a useful distinction, comparable to the one between fixed and variable/marginal costs. The inevitable conclusion within this analytical framework is that in a situation of categorical crowding out, the incentive should be increased in magnitude to offset the crowding out. Had the fine for parents in Haifa been higher, Bowles concludes, the outcome would have been different: smaller delays, or perhaps no delays at all.

[1] U. Gneezy, A. Rustichini, "Pay Enough or Don't Pay at All", *Quarterly Journal of Economics*, 115(2)/2000, 791−810.

With basic analytics provided, Bowles emphasises that "Learning more about the crowding-out is the next challenge for the Legislator" (p. 56). Perhaps it is so, but nonetheless it does apply only to legislators; in addition to the academic world, it is also about the complex mechanism design in sub-statutory texts, bylaws, contracts, even unwritten rules and any policy, public or not, at any level. "Learning more" is basically about two questions: (1) Does crowding out exist, and how strong/widespread is it? (2) What are the mechanisms of causality from incentives to crowding out?

As to the first question, Bowles' answer is massive referring to various experiments and their results. A substantial number of games specified in game theory have been tested experimentally and the results are reported throughout the book: prisoner's game, public goods game with and without punishment, ultimatum game, dictator game, and trust game. All reported results provide some evidence that behaviour cannot be explained by incentives only – at least the theoretically predicted outcome based on incentives alone is different to the experiment results. Nonetheless, these experiments capture only a small number, a minority of situations in economic life, and it is, at least from the information provided in the book, extremely difficult to estimate the frequency of situations in which incentives do not work properly, i.e. situations in which the outcome is different from the expected. Furthermore, such a discrepancy does not provide evidence that crowding out exists; perhaps it is only evidence of inappropriate incentives mechanism design. Finally, the reader receives no information about the intensity of the crowding out effect even if it exists at all.

Nonetheless, methodologically it is more important whether the results of the human behavioural experiments, a cornerstone of behavioural economics, are reliable, whether these results are a good approximation of real-world behaviour of people. Bowles provides devastating criticism of experimental economics in four points: (1) experimental subjects typically know they are under the researcher's microscope, and they behave differently from how they would under total anonymity or under the scrutiny of neighbours, family or workmates; (2) experimental interactions with other subjects are typically anonymous and lack opportunity for ongoing face-to-face communications, unlike real world interactions; (3) subject pools (to date, overwhelmingly students), may be quite different from other populations, due to their age, process of recruitment and self-selection, creating an unrepresentative sample; (4) social interactions that are studied in experiments are not representative for social relations, since they are focused on settings in which social preferences are important, unlikely many other situations, for example most market transactions. Based on these insights, Bowles concludes that "It is impossible to know

whether these four aspects of behavioral experiments bias the results in ways relevant to the question of separability" (p. 71). It is, nonetheless, puzzling that the author, who subscribes to such scepticism about experimental results, painstakingly uses these results throughout the book as key evidence of the absence of separability, that crowding out exists, and because of that he emphasizes that insight even in the title of the book: "good incentives are no substitute for good citizen".

The second question deals with the mechanism of causality from incentives to social preferences. Bowles identifies two causality mechanisms. The first mechanism is situation-dependence of social preferences. It arises because of heterogeneous repertoire of social preferences of individuals – "our preferences are different when we interact with a domineering supervisor, shop, or relate to our neighbors" (p. 85). The second mechanism is that incentives alter the process in which people acquire social preferences over their lifetime – social preferences are endogenous. Both mechanisms are based on the rather simple insight that incentives, since their purpose is to provide information to the target, inform the target, i.e. agent, about the principal who designed them, about his/her beliefs about agent, about the nature of the task to be done, and, above all, about the presumed motives of the principal who created the incentives. In many cases incentives are "bad news" about the principal – "we are no longerfriends" – changing not only the rules of the game, but its very name. The information incentives inevitably produce moral disengagement or activate control aversion. Furthermore, it can induce a switch in the way of responding to stimulus: affective and deliberative, according to dual process theory, a notion borrowed from psychology. Even, according to the neuroscientific evidence, different regions of the human brain are activated: in the case of incentives – the (deliberative) prefrontal cortex, and otherwise the (affective) limbic system.

Notwithstanding how convincing these findings are, the notion of endogenous social preferences is relevant for examination of the long-term trend of changes of values. Hence the crucial question is not whether incentives sway social preferences, but in what direction. Do market incentives create more or less pro-social values and behaviour? The classical authors disagree on this. For Montesquieu "where there is commerce the ways of men are gentle". For Marx, contrary to that, capitalism and market economy "...is the time of general corruption, of universal venality." Bowles applies a simple reality check – it is evident that the countries where capitalism was born, Western and Northern Europe, are not societies of "universal venality" at all, on the contrary. The problem for the author is that, using his own conceptual framework, there is no other conclusion but that market incentives in the crowding-in process, produced pro-social preferences of the population, and there is ample em-

pirical evidence that preferences are more pro-social in the countries with long capitalist tradition than in other societies. And that conclusion does not speak well about crowding out – which is the major topic of the book. As the antidote to this inevitable conclusion, Bowles suggests a hypothesis about the mechanism the crowding-out effects of market incentives that are offset by institutions of liberal society. Alas, there is no answer to the question why liberal society has been established only in market economies, with all incentives that they provide.

Hence, taking all these things into account, how should legislators behave? After explaining, at least up to a point, a rather nebulous notion of the liberal trilemma (the impossibility of Pareto efficiency, preference neutrality, and voluntary participation), Bowles introducesthe concept of the second-best world. All that is a preparation for the firework of Bowles' own insights that are labelled the "five uncomfortable facts about incentives" that legislators ostensibly learn about after "visits to economics faculties". In his own words: "(1) incentives are essential to a well-governed society; (2) incentives cannot singlehandedly implement a fully efficient use of economic resources if people are entirely self-interested and amoral; (3) ethical and other social preferences are therefore essential; (4) unless designed to at least 'do not harm,' incentives may stand in the way of 'creating better people'; and (5) as a result, public policy must be concerned about the nature of individual preferences and the possibility that incentives may affect them adversely" (p. 185).

This deserves a brief comment: (1) true; (2) not true, see First fundamental theorem of welfare economics, (3) ethical and other social preferences cannot be essential if they do not exist and they do not exist based on the assumption in the previous insight, (4) the possibility exists, no doubt, but one should refer to Montesquieu and empirical confirmation of his prediction about the long-term effects on endogenous social preferences; (5) there is no results whatsoever, or at least it has not been demonstrated in the book, so this normative "fact" about public policy is not relevant. The last point deserves more deliberation. For all the insights in the book, from all the theoretical considerations and experimental results, it is evident that social preferences are relevant in specific situations,which includes personal contact, long term relations, many of them being face-to-face relations. Hence it is about effectiveness of incentives created by contracts, agreements, bylaws, unwritten rules, even one-off orders or guidelines. It is about private, not public policy. It is not about legislators, but about all of us and how we build our social relations. If there is a job for legislators in this area at all, it is to create enough free space for individuals to be engaged in the search of their own optimal mechanism design and to voluntarily conclude contracts that will encompass it. Also, to provide enough legal clout for the *pactasunt-servanda* principle not to be violated.

The book effectively ends with advice for the Aristotelian legislator, the one that will make people better, in ten points (p. 206–7). Some of them are intuitive and reasonable ("in the presence of categorical crowding out, avoid small incentives", "avoid moral disengagement", "avoid control aversion"), but it is questionable how legislator can apply them. Some can be counterproductive – Bowles recommends that legislators should avoid creating "perfect conditions necessary for market to work well in the absence of social preferences". The point is that the market is depersonalised, that social preferences, as Bowles himself explains in his book, are not a salient factor in market transactions, so it is puzzling that the advice to legislators is not to improve conditions for market transaction, not to improve incentives to the exchanging parties to be efficient.

At the end of the day, public policy is inevitably about incentives. Some of them are effective, some are not. Some of them are good, directing individuals towards desirable behaviour, some of them are not. The focus should be to the following question: why are some incentives ineffective and bad? The answer would enable principals to select effective and good incentives. Advances in theory of mechanism design (mainly by 2007 Nobel prize winners Leo Hurwicz, Eric Maskin, and Roger Myerspon) promise that the search for good and effective incentives will have theoretical underpinning, rather than being a purely trial and error process. Bowles' book is a significant contribution to the debate about optimal mechanism design related to incentives. If the book is considered this way, it is a valuable achievement.