Can rules and institutions be sustained by self-interested agents?
A theory of diluted incentives, monitoring cascades, and social collapse

Extended Abstract

Bruno Strulovici
Northwestern University

December 30, 2016

This paper is concerned with the following question: can a society populated by purely self-interested agents sustain socially beneficial rules and institutions? The central premise of the paper is that the information necessary to enforce rules and institutions cannot be taken for granted: For example, criminal investigations are difficult tasks which may be botched or biased. Socially-valuable regulations concerning labor, health, or other public matters may be secretly ignored or circumvented, and those in charge of enforcing them may be bribed, deceived, or tempted to abuse their power.

The model developed to conduct the analysis relies on three key assumptions. First, each signal about agents' behavior is either costly to acquire, subject to manipulation, anonymous, or non-verifiable. For example, if a crime takes place, the occurrence of the crime is itself a signal, but an anonymous one in the sense that it does not identify a specific culprit. An investigation may reveal the culprit's identity—a non-anonymous signal—but it is costly to run and subject to evidence fabrication, tampering, or destruction. An individual may claim to have witnessed the crime by chance (and, hence, "for free"), but may in fact be lying, intent to harm the accused or to protect another person, or even cover his own wrongdoing. Second, the population is large and the interactions considered here are disciplined by third-party enforcement. For example, if a firm executive abuses an employee, or if a competitor sabotages a firm's output, the abuse must be investigated and punished through legal means rather than by direct retaliation. Third, the investigation process following a crime or failed outcome has a sequential "monitoring-the-monitor" structure: the case is first assigned to an individual who may conduct a genuine and costly investigation, shirk, fabricate or destroy evidence, and/or be bribed by the person under investigation. The case is then transferred with positive probability to a second individual who may, depending on the first investigation's findings, look into the original crime or into the first investigator's actions, and has the same options to shirk, fabricate evidence, etc. No a priori bound is imposed on the length of this monitoring chain.

The model is otherwise quite general: the set of actions that each agent may take, how they interact with other agents' actions, and the payoffs that result from them, and how these features evolve over time and depend on past outcomes and investigatory findings is left quite arbitrary. Agents' rewards and punishments lie in an arbitrarily large but finite set, and agents are arbitrarily patient. Punishments entail no agency cost and can be administered arbitrarily over time.

Under these assumptions, the model has an essentially unique equilibrium, in which no agent ever makes any effort: agent's failure to internalize the value of institutions and rules leads to a complete social collapse. The model pits standard carrot-and-stick incentives, delivered in the current period or through continuation payoffs, against several forces which, combined together, lead to the collapse. First, agents have many opportunities to "misbehave" (e.g., steal, violate rules, lie, fail to help other agents, etc.). While it is easy to incentive an agent to make an effort on a small set of clearly identified tasks, it becomes arbitrarily difficult to provide such incentives for many possible tasks that the agent may anonymously be faced with. For example, one may make an agent "responsible" for what happens to a given individual (project, firm, etc), or a small set of individuals, but not for what happens to all the individuals that the agent may face on any given day. As a result, a selfish agent given some random opportunity to misbehave may a priori do so with impunity, unless a serious investigation takes place to identify him. This observation is formalized using recent results concerning the statistical properties of large sets of independent random variables.

This leads to the second part of the argument: the impossibility of implementing accurate but costly investigations. Consider, first, an individual in charge of an investigation. To incentivize this individual, one must either punish him if he botches the investigation or reward him if he "solves" the case. The punishment approach clearly requires the introduction of a second monitor. Because the reward approach creates an incentive to fabricate evidence (e.g., appear to solve the case by forcing confessions or testimonies), it also requires the presence of a higher monitoring level to discipline the first investigator. Moreover, both approaches are susceptible to bribes: the subject investigated may pay his investigator in exchange for discarding evidence, which can only be prevented by introducing a higher monitoring level. Of course, the second monitoring level creates the need for a third one, and so on. Notwithstanding the cost of all the entire monitoring structure (which is finite if investigations stop at each level with positive probability), sustaining an *a priori* unbounded monitoring chain is problematic for other reasons. First, along any monitoring chain featuring active investigations at each level, the probability that the agent at the origin of the case is guilty must, conditional on no incriminating evidence being produced during the first n levels, become smaller and smaller as n gets large. This means that a monitor's incentive to seriously re-open the case when no evidence was produced until his level was reached becomes arbitrarily weak and must eventually be dominated by the investigation cost. Second, bribes severely affect the set of incentive schemes that a regulator may implement, by artificially reducing a guilty agent's punishment and increasing a monitor's reward. Consider for instance the value, for the first monitor, of fabricating or manipulating evidence. This entails an immediate reward for (seemingly) solving the case as well as a probability of punishment if the fabrication is uncovered. To offset the fact that the punishment arises with a smaller probability than the reward, its magnitude must be larger. Moreover, if the higher-level monitor who uncovers the fabrication can be bribed, his effective reward must match the effective punishment of the first monitor, both being equal to the exchange of money between the parties. To dissuade evidence fabrication by the first monitor, therefore, the second monitor's effective reward must, under general conditions, strictly exceed the first monitor's reward. Iterating this observation leads to impossibly large rewards for later monitors.

Beyond its robustness, this result raises other issues of interest. In applications where free public signals are available (e.g., athletes facing off in broadcasted events), agents' obedience of the relevant rules is readily observable, free and non-anonymous, and seemingly immune to the problems described here. Even in these cases, however, the mechanisms described here are relevant: for example, they may affect the extent to which doping can be controlled, and suggest the importance of whistleblowers, willing to expose or produce the signals needed to discipline the agents violating the rules at a personal career or financial cost for themselves. Likewise, the public facade of law enforcement may seem reasonably well observable: citizens can see how safe their neighborhood is and exchange information about the behavior of law enforcement officers. Beyond this public facade, however, the higher levels of monitoring are less transparent: bribes of public officials, for example, are not exposed freely: they require costly investigations or risky denunciations by insiders.

With the development of increasingly sophisticated monitoring technologies, one may also be tempted to argue that free, accurate signals will become more easily available. For example, the introduction and proliferation of surveillance cameras in streets, buildings, vehicles, and as part of law-enforcement officers' uniforms make both crimes and intervention decisions easier to monitor.[1] However, besides the obvious risk of abuse occurring if such information falls into malevolent hands and, more generally, the agency problems linked with the proper handling and release of such information, new technologies also create new ways to commit crime, such as hacking into computer systems or databases, spreading libelous rumors, engage in identity theft, use stealthier doping or other law-evading techniques, etc. Moreover, because a key determinant of criminal behavior concerns the intent to commit crime (*mens rea*), in addition to proving commission of the act itself, new technologies may have limited effectiveness in ascertaining the guilt of some defendants; for a recent example, body and dashboard cameras have sometimes failed to convey a clear understanding of the officer's perception of danger. These observations suggest that the

[1] Soccer fans may also ponder the feasibility of Diego Maradona's "hand of God" in the era of new technologies.

mechanisms described here may not soon be dismissed on technological grounds.