10

Challenges for Market and Institutional Design when Countering Exploitation Strategies

Gigi Foster, Paul Frijters, and Ben Greiner

Abstract

Cooperation within larger groups is often endangered by incentives to free ride. One goal of market and institutional design is to create environments in which socially efficient cooperation can be achieved. The main point in this chapter is that only considering first-order incentives to cooperate within a larger group may not be sufficient, as subcoalitions display reciprocal behavior despite the incentives to renege. Three related complications are discussed: (a) exploitative behavior is often coordinated in subgroup coalitions, (b) natural within-group resistance to exploitation already exists, and (c) the actions of group members can often only be imperfectly monitored. Given these realities, implications of current research for applied market and institutional design are outlined.

Introduction

A well-known problem for any social group is that some individuals may find it more rewarding to try to steal the product of the group's activities, instead of making whatever investments are required to join the group and participate legitimately in the group's productive activities. There are many forms of such exploitation. In the human world, they range from free riding on publicly provided goods (i.e., consuming such goods without having paid for them), to untrustworthy behavior in markets and corruption within organizations, to rogue jihadist groups stealing resources in unstable regions. In the animal kingdom, examples of exploitation abound. From birds who brazenly steal other birds' parenting resources (Davies 2000), to insect fathers that pass genetic codes

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

onto their offspring that greatly increase the offspring's chance of rising in the social hierarchy while avoiding detection by competing fathers (Hughes and Boomsma 2008), to bacteria that cheat their peers (McGinty et al. 2011), the social structure of animal species leaves room for individually advantageous cheating features (for more animal examples, see Dubois et al., this volume). Humans, clearly, are not the only animals to have hit upon cheating as a potentially viable survival strategy.

Market design "helps solve problems that existing marketplaces haven't been able to solve naturally" (Roth 2015:7). That is, economists as market designers have started to take over a role which had previously been largely occupied by entrepreneurs and lawmakers by not only trying to understand the working of markets but also using that understanding to rewrite the rules of markets in order to fix them when they are broken. Applied to the context of exploitative behavior, market design (or organizational/institutional design)[1] with a social welfare objective seeks to make anyone's participation in an institution or market safe from being exploited. The lessons learned from this endeavor may also be able to inform research about cooperation within and between other species. In designing ecological environments, for example, humans may be able to achieve states that allow endangered species to survive.

To research the underlying social dilemma, behavioral and experimental economists have modeled exploitation as a public goods game (where it is socially optimal but not individually rational for each individual to contribute to the public good) and examined the effects of such things as group size, group composition, the size of free-riding incentives, the possibilities for leadership, and other aspects of the group environment on the level of contributions (for literature surveys, see Ledyard 1995; Chaudhuri 2011). This is different from how cooperative investment and exploitation is modeled in evolutionary biology. In the latter, the (co)existence of investors and exploiters is the outcome of an evolutionarily stable equilibrium in a model that assumes all individuals to be selfish with respect to increasing their chances of reproduction. In (proximate) behavioral economics models, such as the one outlined below, cooperation (i.e., nonexploitative investment) is a choice that cannot be rationalized when assuming purely selfish individuals maximizing their short-term utility. However, empirically, some individuals deviate from the behaviors predicted for rational selfish agents, even in the absence of institutional interventions. These deviations may result from the incompleteness of economists' models of individually rational behavior, which arguably exclude some individual traits that evolved due to the longer-run dependence of individual survival and reproductive success on the social group structures in which individuals operate. Nonetheless,

[1] In this chapter, we use the terms "market design" and "institutional design" more or less synonymously. While the targets of economic design differ, the approach is the same: creating rules and incentives such that the market designer's objectives (which are here assumed to be aligned with society's objectives), in terms of outcomes, are met.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

our focus here is on the short-term, proximate utility model, to highlight the immediate incentives of individual actors and how to deal with them.

In a typical behavioral economics model, the simplest setup of a situation in which agents may benefit from the investments of other agents can be summarized by the following reward equation. The equation describes the utility payoff of an individual, Y_i, based on the agent's own economic choice, x_i (where a higher value denotes a more prosocial choice), and the choices of others, x_{-i}:

$$Y_i = f_i(\mathrm{x}_i, \mathrm{x}_{-i}). \tag{10.1}$$

A social dilemma emerges when for some individuals

$$\frac{df_i(.,.)}{dx_i} < 0, \text{ and yet } \frac{\sum_j df_j(.,.)}{dx_i} > 0. \tag{10.2}$$

That is, being more prosocial is associated with a loss to the individual but with a net benefit to the whole group (and being less prosocial has the opposite set of effects). The utility function f_i can differ between individuals such that in equilibrium, some individuals may contribute while others free ride.

Market and institutional design can be thought of as the creation of institutions that alter the relationship between individual and group interests (see also Ostrom et al. 1992). If a social welfare-oriented market design is effective, then it aligns the interest of the individual with the interest of the group. To allow for such institutions, we can extend the reward equation with institutional investments α_i and α_{-i} that run from low to high:

$$Y_i = f_i(x_i, x_{-i}, \alpha_i, \alpha_{-i}), \tag{10.3}$$

with $\alpha_i = \alpha_{-i} = 0$ in the case that no institutions exist, resulting in the dilemma situation described above. An effective market or institutional design alters the dilemma game such that it allows for investments α_i that affect the relationship between x_i and Y_i. At some point the total investments $\sum_j \alpha_j > 0$ are large enough such that:

$$\frac{df_i(.,.,\alpha_i,\alpha_{-i})}{dx_i} > 0. \tag{10.4}$$

In this case, it becomes individually rational to contribute to the public good, even for individuals who would free ride if no institution were in place.

The investment α_i may be costly. It could represent a credible commitment to punish, as in the classic public goods games which feature a punishment option (Fehr and Gaechter 2000); in this case, an individual's contribution to punishment represents a contribution to a second-order public good (for discussion, see Yamagishi 1986). If a sufficient number of people plan to punish selfish behavior, and this information is public, then the overall level of

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

punishment expected for not contributing may be sufficient to incentivize efficient contributions to the first-order public good. A large laboratory experimental literature, starting with Fehr and Gaechter (2000), shows that allowing for costly punishment increases cooperative investments in social dilemma situations. Krueger and Mas (2004) and Mas (2008) provide empirical evidence for the relevance of costly punishment outside the laboratory.

The institutional investment α_i may also be the establishment of a reputation system, such as the feedback systems in online markets like eBay. Giving feedback may be costly: it requires taking the time to log in and give honest comments, which if negative may result in retaliation, creating a further cost. If adequately utilized, however, this type of feedback system may be effective in incentivizing buyers and sellers to behave in a trustworthy manner in the online market platform (see also Bolton et al. 2013).

As a general principle, the function of market design is to provide institutions that either affect the relationship between x_i and Y_i directly, or provide an infrastructure such that investments α_i can affect the relationship between x_i and Y_i at a reasonable cost. Several such options might immediately come to mind: allowing communication and increasing visibility, closing loopholes that can be strategically exploited, or providing more effective punishment options.

However, the social world is considerably more complex than the picture painted above. As a result, direct, simple measures that only address first-order dilemma incentives may be ineffective. Merely ensuring that deviations by a single individual are not profitable may be insufficient and, in the worst case, may even create an effect opposite to what was intended.

To illustrate this, we discuss three types of complications:

1. Exploitation is often carried out in *groups* rather than by individuals. This is because individuals intending to exploit others often must cooperate with each other to do so. Effective institutional design will thus need to provide institutions that steer group dynamics toward inhibiting collusion within exploitative subgroups, while at the same time encouraging cooperative investments within the larger group (or "society").
2. Within many social environments, counterforces are already at work to combat social exploitation. In many cases, these counterforces take the form of altruistic punishers, who prosecute exploiters at their own cost and without formal institutional arrangements. Effective institutional design may need to provide support for these altruistic individuals and be careful not to crowd out their motivation and efforts.
3. In real-world environments, information about others' behavior is usually noisy. Such imperfect monitoring may have a significant impact on the effectiveness of punishment, even with small amounts of noise. Effective institutional design needs to take into account the effects of type I errors (i.e., investors being punished as exploiters) and type II

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

errors (i.e., exploiters not being punished) under conditions of imperfect monitoring.

Below, we address each type of complication, discuss relevant recent research, and highlight challenges that arise. We conclude with a discussion of market and institutional design based on preliminary research results, review implications for designing enhanced cooperative investment among nonhuman species, and propose directions for future efforts.

Collusion within Subgroups

The current economics literature on social dilemma games often assumes there is a dominant group from which individuals can only break away alone. However, during our long human lives we live in many different groups in which we are shaped, produce, fight, share, love, reproduce, shape the next generation, and die. These groups range from the family in which we are born to the organizations for which we work, the causes we support, the countries and sports clubs to which we belong, and the organizations and families we set up ourselves. We function simultaneously in many groups, sometimes switch groups when the opportunity arises, and sometimes initiate new groups and subgroups that further our interests (Frijters and Foster 2013:169–170).

The main behaviors that can hamper the efficiency and effectiveness of groups are then not only the free riding of individuals, but also the free riding of subgroups who do not pull their weight: the exploitative behavior of cohesive privileged groups that comes at the expense of society as a whole. A prime economic example of such antisocial subgroup cooperation, often termed "rent seeking" in economics, is seen in cartels. Companies in cartels cooperate in their choice of prices such that they all make greater profits at the expense of the larger society (for examples of conceptually similar behavior exhibited by nonhuman animals, see the discussion about "ganging up" in Dubois et al., this volume). Other examples within countries are criminal gangs or the regulatory capture of, say, financial watchdogs by financial interest groups that use that capture to evade monitoring by society as a whole. An example of exploitative coalition formation on the international scale is the phenomenon of individual countries (which are subgroups of "the international community") reneging on agreements about global climate change.

One puzzle here is how these groups manage simultaneously to cooperate within their subgroup and selfishly exploit the larger group. That is, the individuals in these groups behave at the same time selfishly (with respect to the larger group) and in a trustworthy fashion (with respect to their fellow coalition members). How can you trust someone who is observed to cheat someone else at the very same time? When it comes to whole countries breaking away from international agreements, the answer is simple: the prime loyalty of individuals

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

is toward their country, not toward an international agreement. Within countries and especially within individual markets, the answer is less clear.

The concept of "group identity" (Akerlof and Kranton 2005) may be helpful in this context. Differences in behavior toward a subgroup and the larger group (or other subgroups) may be sourced in a higher identification with the smaller group which in turn translates into a higher utility derived from adhering to that smaller group's norms, as opposed to universal norms. It may even trigger stronger punishment within the smaller group, enforcing these norms (Goette et al. 2006).

From a market design perspective, the question becomes how to disincentivize cooperation within an exploitative subgroup to increase cooperative investment within the larger group. With respect to cartels, for example, the literature on competition policy already proposes some methods of combating such collusion: closer supervision, preventing within-subgroup communication, whistle-blower provisions, and so on. Essentially, the existing identity with, and loyalty to, the larger group is harnessed and directed toward establishing and implementing monitoring and punishment mechanisms against any would-be smaller group.

The simple mathematical framework outlined above can be used to model the formation of a subgroup within a larger group such that cooperation within the subgroup occurs at the expense of the production and consumption of the larger group. This conceptualization of the problem accommodates many forms of corruption and collusion. Individuals face the option of being prosocial within a clique, but antisocial toward the group as a whole. There is thus a choice c_{ij}, denoting an action taken by i that affects j (where j is the individual or subgroup favored), which is nonetheless antisocial from the point of view of the group as a whole. The individual or subgroup j can then reciprocate this choice by choosing an appropriate c_{ji}. In terms of the effects of these choices on the individual's material rewards Y_i, we would have the following situation:

$$\frac{dY_i}{dc_{ij}} < 0, \frac{dY_i}{dc_{ij}} + \frac{dY_i}{dc_{ji}} > 0, \frac{dY_j}{dc_{ij}} + \frac{dY_j}{dc_{ji}} > 0, \sum_k \left\{ \frac{dY_k}{dc_{ij}} + \frac{dY_k}{dc_{ji}} \right\} < 0. \quad (10.5)$$

Expressed in words, these four conditions say that the antisocial choice c_{ij} costs the individual (first condition), that the individual nonetheless gains if the other individual or clique j reciprocates (second condition), that the individual or clique j gains from the antisocial choice and its reciprocation (third condition), and that there is still a net cost to the whole set of individuals from the antisocial choices and their reciprocation (fourth condition). The question is: What institutions can be devised to break the "benefit" to cliques (i and j) of being antisocial from the point of view of the group as a whole?

Two recent studies provide some initial insights into the endogenous formation of exploitative subgroups. In an ongoing study conducted by two authors of this chapter (Foster and Greiner), human subjects are invited to an

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

experimental computer laboratory and take part in a game in which they can earn a substantial amount of cash.[2] Participants interact over 100 rounds in fixed groups. In the first stage of each round, subjects vote over who will get to allocate group resources (a monetary amount) as a dictator in the second stage. This setting is extreme in the sense that the game has an "empty core": cooperative game theory predicts that in this game there is no stable coalition of a subgroup of players that would be robust with respect to counteroffers made by other players. (Given small voting costs, noncooperative game theory, in fact, predicts that every member will vote for him/herself.) The experiment involves three possible treatments:

1. There is no communication before the dictator election takes place.
2. There is a stage in which each group member submits a nonbinding distribution proposal (similar to a cheap-talk election campaign).
3. Proposals are submitted and are binding, such that group members vote for a distribution as proposed by one of their members.

Results from this study show very strong and long-lasting coalitions, the size of which depends on the institutional setting. In particular, when there are no campaigns or only nonbinding proposals, in many cases a minimum majority (i.e., an exploitative coalition containing three out of the five subjects) was observed among whom the elected dictator distributed all resources, to the exclusion of others. Nonbinding counterproposals made by the excluded subjects trying to break up these cliques were often unsuccessful. Only when proposals were binding did groups often end up in so-called "grand coalitions," featuring fair distributions of resources among all group members.

In a way, these results mimic the emergence of clans or families within larger societies. The treatment effects suggest that if market design can affect the informational environment, then it also can affect the emergence and stability of exploitative coalitions. With limited communication between group members, only the current resource owner can communicate credibly through an actual allocation decision. Even if the institutional setup allows for cheap-talk communication, the current dictator has the advantage of being able to support his/her own proposals with evidence, while counterproposals from others in the group lack this credibility. However, when individuals can credibly commit to proposals, there is a lower likelihood that an exploitative subgroup grabs all resources and distributes them among themselves, and a higher probability of the emergence of a fair and equal distribution.

Foster and Greiner make another interesting observation in their experiment. In the treatment with nonbinding proposals, two out of three experiment sessions had to be stopped prematurely, at rounds 50 and 60, respectively. This was because some subjects in these experiments became very upset about being, in effect, permanently excluded from the group's resources, even though

[2] A first version of the research paper on this study is expected to be published by the end of 2016.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

they made generous counteroffers. In looking for possibilities to punish the exploitative behavior that victimized them, they found their own way of breaking the rules and starting a revolution: they delayed their decisions—for which a time limit was not enforced in the experiment—such that the experimental session progressed at an increasingly slow pace and eventually had to be prematurely stopped.[3]

The second study, by Murray et al. (2015), provides evidence about how inefficient coalition formation arises. Here, fairly small groups of subjects (four or six) played a game in which whoever was the leader in a round had to choose a partner in that round who would get a large bonus and who would also be the sole producer for the whole group. Those who were not chosen only received a share of this group production, but not the relatively large additional bonus of being the producer. In a subsequent round, the former partner became the leader, or had a high probability of becoming the leader, which endogenously gave rise to long-lasting partnerships that thrived over many rounds, at the expense of the group as a whole. An overall loss in welfare came about from the fact that the productivity of each person varied in each round, meaning that these sticky partnerships prevented optimal allocation—in which the chosen partner would always be the most productive person in that round—from emerging. The experiment featured the type of back-scratching behavior observed inside major institutions and within clans and other larger groups, a major problem in both developed and developing countries (e.g., Murphy et al. 1991).

The experiments of Murray et al. (2015) included the introduction of market design institutions aimed at breaking up inefficient partnerships once they had formed, by either reducing the bonus that the leader would allocate or by randomly breaking up a coalition and putting someone else in charge as leader of that round. It turned out to be very difficult to split up coalitions: the dominant pattern was that former partners who were randomly broken up would simply reestablish their coalition as soon as they were in the position to do so, often leading to subgroup coalition formation by everyone in the experiment: that is, the whole group became divided into a set of (active or dormant) dyadic partnerships. Even when bonuses became low enough that even the coalition members would be better off without being in a coalition, they largely endured, perhaps due to fear that others would otherwise come to dominate the leadership positions.

So far the main implication from this line of research is that some institutions might be advocated to prevent subcoalitions from forming, while other institutions might effectively dismantle them. In particular, juxtaposing people who are unfamiliar to each other appears to help delay coalition formation, and this can be combined with setting low levels of discretion (high levels

[3] To us, as experimenters, this was actually bad news, since we lost experimental control (in the sense of defining and controlling the available strategy set). Nonetheless, it is an important observation.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

of monitoring), an institutional backdrop that lowers the benefits of coalition formation. Human warfare provides classic examples of attempts that make it hard for soldiers of opposing armies to fraternize or desert through monitoring. In the eighteenth century, for instance, Frederick the Great used very conspicuous uniforms for several of his troops, so that they would be (a) clearly visible to the enemy and (b) easily spotted should they attempt to cooperate or desert (von Clausewitz 1832/1968:415). Breaking up established coalitions, however, appears to be more difficult.

Existing Counterforces to Selfish Behavior

In many dilemma situations, there exists some spontaneous disciplining behavior. Even in the absence of formal institutions, some people not only invest in the joint group interest (e.g., the public good), but also act as disciplining agents toward those who do not invest. This is true even in nonhuman social animals, in whose societies evolutionary biologists have observed both antisocial behavior and punishment of that behavior, with both explained as the result of genetic competition. As Hughes and Boomsma (2008:5152) state: "The nonidentical reproductive interests of group members inevitably result in individual-level selection favoring cheating and the antagonistic coevolution of cheat suppression."

Casual observation and the results of economic experiments show that this cheat suppression can occur even when disciplining actions are costly. In the background, cheat suppressors are often sustained by embedded social norms, like an individual conscience, that support their behavior. As noted by Frijters and Foster (2013:178) with respect to the enforcement power of such individuals within large groups, even a small number of individuals who are willing to mete out punishment to those who are supposed to be punished might be sufficient to dissuade exploitation by others in the group. Hilbe and Sigmund (2010) and dos Santos et al. (2011) show that when past behavior can be observed, and thus a punishment reputation can be built, the existence of (seemingly) altruistic punishers can be an equilibrium outcome in an evolutionary "meta-game" with purely selfish agents.

Various studies on norm enforcement, starting with Fehr and Gaechter (2000), have found evidence for altruistic punishment. In a typical public good experiment with peer punishment possibilities, punishment is costly, both in terms of the direct costs of executing punishment (in the typical experiment, it costs 1 unit to inflict a punishment of 3 units) and in terms of counterpunishment (often people who get punished then punish back, as a reciprocal action or to discourage further punishment). Altruistic punishment behavior is thought to be driven by an emotional response to norm violations that is acted upon even at a cost to oneself. It is more likely to be observed among those who cooperate most themselves (e.g., Fehr and Gaechter 2002).

From "Investors and Exploiters in Ecology and Economics: Principles and Applications,"
Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

In the language of our framework, altruistic punishers have a utility function $Y_i = f_i(x_i, x_{-i}, \alpha_i, \alpha_{i-i})$ such that $dY_i/d\alpha_{-i} > 0$, meaning that the costs of contributing to the punishment institution are more than counterbalanced by the value derived from such a contribution (e.g., through a positive internal sense of "doing the right thing"). A challenge for market design is then to ensure that the prosocial incentives of these altruistic punishers are not crowded out by introduced institutions.

There is evidence that norm enforcers exist in other environments. Bolton et al. (2015) study the interaction of reputation systems and conflict resolution systems and provide a case study of eBay's feedback system and the feedback withdrawal process. On eBay's pre-2007 platform design, after each transaction, both the buyer and seller could give positive, neutral, negative, or no feedback on the trading partner. Feedback was immediately published on the platform. The feedback could only be withdrawn if both transaction partners mutually agreed to a withdrawal. Such a withdrawal process was introduced to the platform to facilitate the resolution of a conflict after it had arisen, through the "making good" of initial offenses and subsequent removal of negative feedback. However, the withdrawal option also provided strategic incentives that actually may have escalated the conflict in the first place. In fact, the possibility of mutual feedback withdrawal makes it a dominant strategy to respond to negative feedback with one's own retaliatory negative feedback, independent of the transaction details. This is because the retaliation is a low-cost means of creating negotiating power that can be applied during negotiations for mutual feedback withdrawal, which may then occur without the original offender incurring the (presumably higher) cost of "making good."

Using field and laboratory data, Bolton et al. (2015) showed that the existence of a withdrawal option may, in this way, hamper trustworthiness, and thereby trade efficiency, on eBay's platform—as well as lowering the information content of feedback. However, both field data and laboratory data also show evidence of costly altruistic punishment. In the field data, a withdrawal request strategically supported with feedback retaliation (such that in theory, both parties face incentives to withdraw) is no more successful than a withdrawal request that is not backed by this threat. In the laboratory, cooperators emerge who, after receiving negative feedback that is purely strategically motivated, are likely to be much more emotionally distressed and are more likely to punish offending sellers' attempts to enforce feedback withdrawal, insisting instead on the sellers "making good." Strategic retaliatory claims hence appear ineffective, at least when applied to this group of cooperators, which in turn supports the efficiency of the conflict resolution system.

Similar costly punishment can be observed in many social settings, ranging from open disapproval of people who do not tip waiters or who do not wait patiently in a queue, to consumer boycotts of companies that use child labor or flout environmental norms.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

Other studies have found evidence for crowding-out effects. Falk and Kosfeld (2006) found that tightening contractual obligations to cooperate may actually lower overall cooperation. Mellström and Johannesson (2008) studied whether providing financial incentives crowds out participation in blood donation and found evidence of partial support for this conjecture.

Our proposal is that market design efforts to change institutional settings in a group, with the goal of lowering exploitation, should be applied in such a way as to avoid crowding out the incentives of the group's preexisting counterforces to exploitative behavior.

What are the relevant questions to ask if one wished to pursue this proposal? We suggest that they include the following: Who has information about exploiters? How costly is direct punishment implemented by those with this information? And, more subtly, does the social norm being violated by these so-called exploiters truly benefit the group as a whole?

If the information about exploiters is only known locally and is not verifiable at a more aggregated level, and if the cost of punishment is low, then spontaneous punishment via a social norm is probably more efficient than a formal institution. When the social norm itself is not efficient, which can occur if a group's belief about what constitutes the social good is simply wrong, then there would seem to be a role for a group information institution that is able to assess claims about the benefit of this or that behavior in which members may or may not be engaged. For example, such an institution could verify not only whether a company truly has used child labor, but also whether that is indeed a bad thing to do given the actual circumstances faced by the employed children (e.g., if the alternative is child prostitution, then child labor may be the better option). Standard economic arguments on efficiency and transaction costs would apply to this design problem, inasmuch as the role of institutions would be to provide group mechanisms for information dissemination and punishment only when there is a returns-to-scale argument supporting such a role. By contrast, when there are decreasing returns to scale, such as when information is local and nonverifiable, the role of institutions would be more to give official group approval to low-level punishment rather than hindering or channeling it.

Imperfect Monitoring

Most of the current literature on the effectiveness of institutional settings in combating exploitative behavior assumes that each agent can perfectly observe the actions of every other agent. Yet the existence of perfect information in this area, like the absence of frictions elsewhere in an economy, is unrealistic (for further discussion, see Frijters and Foster 2013:70). In the real world, actions are not perfectly observed, and observing the consequences of actions does not always lead to straightforward conclusions about the underlying actions themselves or the intentions behind them.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

To formalize this argument, we can think of x_{-i}, the actions of other players in

$$Y_i = f_i\left(x_i, x_{-i}, \alpha_i, \alpha_{-i}\right), \tag{10.6}$$

as being only imperfectly observed by agent *i*. All that is outwardly observed is a signal s_j drawn from a distribution around each true x_j. Punishment through institutions as well as the determination of future behavior in response to current experiences will then be based on s_j rather than on x_j. This may lead to errors—specifically, the punishment of investors (type I) or the nonpunishment of noninvestors (type II). The risk of either error may result in less prosocial behavior.

Ambrus and Greiner (2012) and Grechenig et al. (2010) tested the effect of noise in public good environments experimentally. In Ambrus and Greiner's 2012 experiment, decisions about whether to invest in the public good are binary (as in classic prisoner dilemma games), and noise is introduced as pure type I errors, through setting a 10% chance that any given investment is publicly shown as a noninvestment. This small amount of noise has a stark effect: investments and payoffs significantly decline and observed punishment increases. The reason seems to be that while group members punish the same way as they do in perfect monitoring conditions, punished investors react adversely and reduce their investments in the next round. This sets off a cycle which ends, after a while, in no investments in the public good by any group member. Essentially, the monitoring problem directly results in the crowding out of prosocial behavior. Grechenig et al. (2010) obtain very similar results, in an environment in which investment decisions are continuous and noise is added as a random shock on the investment of a group member.

Thus, imperfections in monitoring other group members pose a challenge to any attempt to sustain cooperation. Again, the logic sketched above would seem to apply: How might we design institutions such that punishment occurs at the level of those who have the best information?

An additional element comes to the fore in the presence of imperfect monitoring, which is the role of open examples. If monitoring is very difficult, the possibility arises to invest a lot of resources to get clarity on a small set of individual possible cases of exploitation, and then to have excessive punishment for small transgressions if the uncertainty is resolved. This design strategy again reflects basic economics: when it is hard to observe wrongdoing and hence when the odds of verifiable detection are low, the punishment of that wrongdoing when detected should be extreme to provide incentives to cooperate. A good example of this is the social demand made of politicians and judges to be beyond reproach: even small acts of criminality (say, stealing five dollars) would cost them their jobs, which are worth millions (for further discussion of the mechanisms that support this type of social dynamic, see Foster and Frijters 2016).

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

Conclusions and Frontiers for Market and Institutional Design

The three complications of the dilemma of exploitation that we have discussed are not the only ones, yet they are important and nontrivial to address. A market designer attempting to improve cooperation in a real-world dilemma situation will almost surely encounter these issues, and hence we propose that their study is worthy of market design economists.

The fact that exploitation is often carried out through the activities of subgroups poses the problem that market and organizational interventions targeted at a subgroup may also affect the whole group, and vice versa, both in terms of supporting and preventing cooperation. Reducing communication possibilities for firms in a market, for example, may be effective in preventing collusion, but may also reduce overall market transparency and thus lead to a second-best outcome in terms of market efficiency. Similarly, establishing a well-functioning mobile phone infrastructure in a developing country does not only support economic development, it may also improve coordination among militia groups. These examples as well as the results of Foster and Greiner's study suggest that communication is a key factor in within-group cooperation and the creation of an elite coalition. The ability to manipulate communication channels and information flows selectively for different subgroups is arguably a very important tool for an institutional designer.

In addition, an advanced understanding of the inner workings of group decision making will be important. Decision-making settings, such as whether everyone has a voice or whether majority vote determines the outcome, will have implications for institutional designers: What is the best way to sabotage cooperation within rogue subgroups to cause them to break up? In a setting that leaves it to group members to invent freely a means of making decisions, Ambrus et al. (2015) found that in a social decision-making context, extreme opinions are suppressed, and only intermediate and moderate individual opinions have a significant impact on the group's decision. Moreover, in decision-making settings that feature uncertain returns, the most risk-averse individual also has influence. This suggests that majority voting (which implies that only the median opinion in a group matters for a decision outcome) is not necessarily the best model of how groups make decisions; however, changing only the opinions of extreme members may have limited effects on the group's (pro- or antisocial) decisions.

The literature on cartels provides other hints about optimal structures to prevent or stem exploitative behavior by subgroups. Whistle-blower regulations allow members of a cartel to go free when they provide evidence about illegal behavior. In a similar way, other subgroups that exploit the larger group might be broken up by providing incentives to individual group members to stop cooperating within the exploitative subgroup. Our current thinking is that preventing coalition formation will probably be much easier than breaking up established coalitions and for this reason, new institutions and markets should

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

be populated as much as possible with individuals who are not already in coalition with others and have limited incentives to form such coalitions.

Existing, endogenously evolved resistance against exploitation (e.g., in the form of altruistic punishers) may need active support through market design settings to be most effective. For example, reducing the costs of punishment or increasing its effectiveness may lead to a more powerful threat against antisocial actions, and thus to more cooperation and less eventually executed punishment (for related results, see Ambrus and Greiner 2012). On the other hand, the introduction of a police force, for example, might crowd out the motivation of effective vigilantes, and thus lead to a less-preferred social outcome if the vigilante organization was more effective than the police (e.g., due to local information advantages). An institutional designer needs to take such potential effects into account. As a general rule, the key issue is whether there are returns to scale in the aggregation of monitoring and punishment, or whether information and punishment are most efficiently collected and implemented, respectively, at the local level.

Finally, imperfect monitoring poses problems of its own. When punishment opportunities are provided to a social group but group members' actions cannot be observed perfectly, then the possibility of type I and type II errors in punishment arises. As Ambrus and Greiner (2012) show, unfairly punishing contributors may initiate group dynamics that lead to a decline in overall cooperation. Institutional design must address these issues. One way of doing this might be to allow the aggregation of noisy signals across group members, if this improves the quality of the information used in decisions about punishment. Another option is to increase punishment for small transgressions, or to put a lot of effort into reducing the uncertainty around a small set of suggested cases.

We close our discussion by asking what can be learned from this line of inquiry in regard to cooperation in other species and implications for "ecological design." As a preamble, we should mention that the economic concepts of free riding and contributing are not quite the same as the concepts of scrounging and production in animal studies, where the typical producer discovers new feeding grounds and the scrounger tags along and eats part of the discovered resource. Only when scrounging is seen as a persistent tactic and the scrounger and producer are not a genetic team (i.e., close kin) would it fit our formulation that in a utility sense the scrounger is free riding and the producers are contributing. When individual animals take turns at scrounging, the situation may in fact be an optimal form of sharing information and highly cooperative. Our comments below, therefore, refer to persistent behavior of an individual or (sub)group, rather than efficient specialization.

Much of the literature in evolutionary biology and animal social behavior takes it as given that both cooperation and cheating will be observed in many social group settings. The economist's singularly interventionist reaction to this reality can potentially bring new insights to the "ecological design" endeavors

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

of those interested in changing the balance of cooperative and exploitative behavior in animal groups. In some cases, the exploitative behavior observed in animals may be seen as bad for humans; for example, when it threatens the survival of whole groups (species) that are endangered or otherwise already stressed. In other cases, humans may wish to augment the naturally occurring exploitation in competing species (e.g., as a means of fighting the spread of undesirable microbial colonies) by pushing them toward self-destruction through the selection of ecological design settings which encourage antisocial behavior within these groups.

Altruistic behavior in many animal species has been found in a number of studies to be genetically coded (e.g., Giraud et al. 2002; Dimitriu et al. 2014), although genetic coding of options may not be required in animals provided with brains that make decisions. Indeed, as pointed out by Dubois et al. (this volume), behavioral ecologists adopt the "behavioral gambit" which assumes that decisions made by nervous systems replicate the evolution of genetic alternatives. With one exception, discussed below, a genetic basis for human altruism is not (yet) a finding, and is not often even hypothesized, in the social sciences; rather, interpersonal differences in levels of cooperation within human groups are typically seen as the outgrowth of differences across people in their cultural programming or in the personal incentives they face. In line with this perspective, our market design approach to supporting cooperation disregards the extent to which human cooperation may be (at least to some extent) genetically coded, since such coding cannot realistically be manipulated or even observed by benevolent designers of institutions intended to promote cooperation. Instead, we focus on those differences in cooperation, either across or within groups, that arise due to differing individual or group incentives; that is, where individuals show *plasticity* in their behavior conditional on circumstances of the environment. It is those incentives and circumstances that are most effectively altered via strategic market design.

The one exception mentioned above is the hypothesis that genetic kinship may increase cooperation among individual groups or subgroups due to basic evolutionary concerns. This "kin selection" hypothesis, originally suggested in the biological sciences (see Wilson 2005), has been tested in human groups and has found some support in research into animal cooperation, such as the cooperation of males in polyandrous tamarin groups who share the care of babies that might not be theirs (Díaz-Muñoz 2011). However, kin selection is debated even in the biological sciences (e.g., see Mathot and Giraldeau 2010). Our concern here is not with genetic sources of cooperation—expressed via kin selection or in other ways—but with cooperation that is manipulable due to the opportunistic capabilities of the species in which it occurs.

In research programs aimed at slowing the growth of bacterial colonies, for example, ecological designers have thought about how to slow down the spread of cooperative genes (as in Dimitriu et al. 2014) or to provide survival advantages to antisocial genes. Our work suggests potential value in also

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

considering how to crowd out cooperative behavior by individual organisms, change the level of noise in signals about socially relevant behavior, and/or encourage the formation of subgroups that cooperate only within themselves. One might achieve the crowding out of prosocial behavior in species that respond prosocially to cues about ambient levels of cooperation by manipulating the signals about levels of cooperation from surrounding individuals (a possibility implicitly suggested by Allen et al. 2016). In other species, crowd out might instead be achieved through the design of local environments in which public goods are already provided in a metered fashion, and hence where the incentives for individuals to produce them cooperatively to ensure survival are reduced. The logic of this approach is illustrated in the health example (Thomas et al. 2012) wherein cancer cells are triggered to become more aggressive (i.e., to exhibit more prosocial behavior, threatening the host) when they are deprived of oxygen—an input necessary for survival, if not exactly a public good. In humans, the level of crowding out of prosocial behavior and the formation of subgroup coalitions are manipulable in the short run due to the adaptability of the human species to environments with different information or norms. To the extent that nonhuman animals are able to adapt to changing environments (termed "plasticity of behavior" in the biological sciences), there may be scope for the discovery of particular mechanisms, such as those suggested above, to achieve short-term changes in crowd out or subgroup formation in such species. In the long run, genetic selection—whether engineered in the lab or invoked in the field through ecological manipulation—is an alternative strategy.

In mammals, coalition formation in response to circumstance has been observed (for a review, see Johnstone and Dugatkin 2000), and scientists have developed models of why these coalitions arise under certain ecological conditions (for a review, see Mesterton-Gibbons et al. 2011). This suggests that in some animal colonies, it may be possible to alter the costs of altruistic punishment and/or the likelihood of subgroup formation in the short run via ecological design. For example, Dugatkin's model suggests that the incidence of coalition formation in primates can be manipulated via changing the amount of resources under competition, the degree of credible exclusion of losers from those resources, and the level of individual investment required to create the coalition (Dugatkin 1998). Such manipulation might in practice consist of altering the distribution of introduced resources among group members; for example, by withholding resources from would-be coalition formers or providing them to individuals in possession of good-quality local information. Other levers that ecological designers might consider, guided by the economic approach, include breaking up existing coalitions through the temporary removal of key members of elite subgroups, or manipulating signal quality—perhaps via the installation of dummies or distractions into the environment.

From "Investors and Exploiters in Ecology and Economics: Principles and Applications," Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21, series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.

Acknowledgments

This chapter greatly benefitted from comments by Sam Brown, Luc-Alain Giraldeau, Kiryl Khalmetski, Julia Lupp, Claus Wedekind, Bruce Winterhalder, and discussions with participants at the 2015 Ernst Strüngmann Forum on "Evolutionary and Economic Strategies for Benefitting from Other Agents' Investments."

From "Investors and Exploiters in Ecology and Economics: Principles and Applications,"
Luc-Alain Giraldeau, Philipp Heeb, and Michael Kosfeld, eds. 2017. Strüngmann Forum Reports, vol. 21,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03612-2.