Local Control: An Educational Model of Private Enforcement of Public Rules

STEFFEN HUCK
Humboldt University[*]

MICHAEL KOSFELD
Tilburg University[‡]

October 1998

Abstract

We study a society of agents where individual incentives conflict with collective ones and thus individual utility maximization leads to inefficient outcomes. We assume that there is no functioning central institution which can control individual behavior. Instead, we analyze a system of what we call *local control (LC)*, where the enforcement of punishment lies in the hands of individuals in the society rather than in the hand of a central institution. The mechanism that governs the spread of control is the educational impact on an agent being controlled by some other agent, where we distinguish between executed and threatened punishment. Agents maximize their payoffs and underlie a constant drift towards not controlling others anymore. Our main results show that LC can survive if the educational impact of control is strong enough relative to the drift. If the educational impact of control is too weak LC breaks down. Moreover, there exists a non–monotonic punishment effect that sets a trap for standard legal policy advices.

Journal of Economic Literature Classification Number: C72, K42

Keywords: cooperation, prisoner's dilemma, social control, punishment, institutional change.

[*] Department of Economics, Humboldt University Berlin, Germany, Email: huck@wiwi.hu-berlin.de.

[‡] CentER & Department of Econometrics, Tilburg University, The Netherlands, Email: M.Kosfeld@kub.nl.

"In order to bring about the transition from present circumstances to those which have been planned, every reform should be allowed to proceed as much as possible from men's minds and thoughts."

— Wilhelm von Humboldt (1792)

"[Punishment] does not serve, or serves only incidentally, to correct the guilty person or to scare off any possible imitators. Its real function is to maintain inviolate the cohesion of society by sustaining the common consciousness in all its vigour."

— Émile Durkheim (1893)

1 Introduction

The question how to restrict liberty in order to protect human beings against each other is nearly as old as mankind itself. It has driven the works of Thomas Hobbes, Jean Bodin, Immanuel Kant and many others. But, phrased differently, it has also been and continues to be a central question of economic analysis: how to achieve efficiency. In general, there may be many obstacles to reach efficiency, but the most important one is, somewhat ironically, economic rationality itself. If individuals behave opportunistically to maximize their income, they will ever so often end up in miserable, Pareto dominated states. This imposes a scientific task going beyond the mere understanding of what happens in the real world: It imposes a problem of social engineering.

Economics, as part of the social sciences, has a successful history in this matter. Most prominently, economic theory has provided many insights in how to create markets such that desirable outcomes are achieved.[1]

But properly designed markets do not always bring salvation. This is especially apparent when interaction is *local*, i.e. when individuals interact only with a small number of others, e.g. with their neighbors. Despite increased mobility local interaction plays still a prominent role in our lives. So, can economists offer advice in how to design institutions when the number of individuals interacting with each other is small? The answer is, of course, it can. In fact, there is a huge body of literature dealing with such issues, e.g. most of principle–agent theory is devoted to this very question.

[1] Recent examples where immediate application followed on theoretical work are the PCS spectrum auctions (e.g., McMillan (1994)) and Alvin Roth's redesign of the market for young American physicians (e.g., Roth (1990) and Roth and Peranson (1997)).

While principle–agent theory or mechanism design usually confines itself to the analysis of bi- or multilateral relationships, the emphasis of this paper is on the society level. In our model, which is of an illustrative nature, there is a large population of individuals each of which interacts with two neighbors. We model this interaction in a way readers will be acquainted with, namely as a Prisoners' Dilemma (PD). The PD is arguably the simplest and most widespread illustration of the earlier mentioned problem that rationality itself may be an obstacle to efficiency. In this game there are two individuals who may either cooperate or defect. Regardless of what the other does, defection always pays individually. Thus, rationality leads to both players defecting which yields payoffs Pareto dominated by the payoffs associated with mutual cooperation.

There are countless studies showing that certain alterations of a standard one–shot PD render cooperation rational — hereby claiming to provide explanations of real–life cooperation (see e.g. the seminal article by Kreps, Milgrom, Roberts, and Wilson (1982)). The present study pursues a different goal as it takes for granted that humans do not always cooperate and, therefore, focuses on the social engineering aspect.

In this paper we are not interested in the possibilities arising when there is a central institution with unlimited power to exert social control. If there was one, the obvious (and trivial) solution to the engineering problem would be the enforcement of cooperation by changing the payoffs in case of defection, e.g. by imposing taxes or introducing punishments. Then, individuals would no longer play a PD but rather a different game where the unique equilibrium has the desirable quality of being efficient.[2] Instead, we analyse a situation where there is only a *weak* central institution which can only act once deviant behavior has been reported by citizens. Only, in that case it can punish deviant behavior. Thus, the model reflects a world which may be characterized by *private enforcement of public rules.* This situation is not completely unlikely. In a recent article, e.g., Hay and Shleifer (1998) discuss the question of legal reform in today's Russia. The authors come to the conclusion that in absence of a well–functioning legal system a promising short and medium term policy is to create public rules which can be privately enforced. In such a world the engineering problem seems to collapse to the seemingly simple optimal adjustment of punishments, but as we will see below this adjustment problem turns out be rather non–trivial.

As agents interact locally and as only neighbours have the opportunity to report deviant behavior to the authorities (or, in short, to punish or control them) we shall speak of a model of *local control (LC).* However, not all individuals will use the opportunity to control or punish as the act of doing so is costly. Rather, we assume that there are two types of individuals — types who will never punish (because they are rational and simply aim at maximizing short–run payoffs) and types who will always punish (because they are by some internal instance driven

[2] Okada (1997) analyzes a situation where the decision for the creation of a central institution is endogenized.

to do so). The latter types shall be called controllers. Both types are assumed to be rational in the standard sense when it comes to playing the game, i.e. they will defect whenever this is money–maximizing.

Now one could argue that people punishing others for defection should also have some intrinsic motivation not to defect themselves. This may be right. But making this additional assumption would mean making life easier for the (theoretical) survival of cooperation, i.e. in certain cases there would be cooperation in the model when there would be none without the additional assumption. But this is a severe problem from the social engineering point of view. For a social engineer it is of not much help to know sufficient conditions for his proposals to work if he cannot take them for granted. This is the same in the engineering of, say, motor cars where an engineer proposing a design for a car that will drive if there is no friction between its wheels and the road would be of little help. Accordingly, we want to make the survival of decentralized social control as difficult as possible.

A further key assumption we make is that individuals may change their type and for the same reason we assume rationality when it comes to playing the game we assume that there is a constant drift away from the social controller. On the other hand, a type who does not punish may become a controller if and only if he is confronted by punishment himself. Here we distinguish between *threatened punishment* and *executed punishment*.[3] This distinction will be of crucial difference. In fact, we will show that the solution to the engineering problem will depend heavily on which of the two educational forces is stronger. Educational effects of punishments are often discussed as indirect deterrence. Salem and Bowers (1970), for example, provide early evidence that formal sanctions can reinforce informal social norms inducing disapproval of deviant behavior and, thus, yielding more compliance. On a more general level, educational effects of punishments require that punishments do not only have consequences but also meanings which can be understood by people. This "expressive dimension" of punishments is discussed, for example, in Kahan (1996) who also provides empirical evidence.

The possibility of type changes through either education or a drift back is the reason why the optimal adjustment of punishments is not as simple as it seems. The severer the punishment, the better the deterrence, one might think. Probably, this is the first intuition of anybody familiar with the basics of economic reasoning in law. However, this intuition may go wrong. In fact, we will show that when executed punishments have a stronger educational effect than the mere threat of punishments, the amount of punishment has a non–monotonic impact on behavior. This effect is so strong that when punishments exceed a certain critical level, the system will experience a total breakdown and will converge to a state in which all players defect all the time.

[3] A similar distinction is common in the deterrence literature where the former's equivalent is referred to as general deterrence and the latter's as special deterrence (see e.g. Beyleveld 1980).

To the best of our knowledge, the theoretical possibility of such non–monotonic effects has been discovered by Akerlof and Dickens (1982) who show that these effects can arise when individuals experience cognitive dissonance after decisions.[4] This is in so far related to our model as cognitive dissonance does something similar in their approach as education does in ours: It changes individual evaluation of outcomes. In a certain way one could say, it changes their preferences and, while with common stable preferences incentives always have a monotonic effect, this does not hold when they can change.

On a less applied level we like to claim that our model adds to the understanding of how social control can work without a powerful central institution. First of all, it does not work always but may require a lot of fine tuning and if, for example, punishments are difficult to adjust (because they are not executed by a legal system but purely rely on social sanctions) it may not work at all. On the other hand, we can see some important ingredients which may make it work, the most important one being local interaction. Local interaction makes it possible to ascribe bad things happening to the action of individuals, it makes them identifiable and this is the prerequisite for calculated punishment (in a broad sense). The assumption of local interaction seems, as already implicitly mentioned above, well–justified. It is not an artificial assumption but rather captures many close–knit aspects which are still important in our lives (see e.g. Ellickson (1991)).[5]

In a recent paper Sethi and Somanathan (1997) discuss the evolution of social norms within a situation of common property resource use, a question that is quite similar to our analysis of decentralized social control. However, our paper differs from their approach in three main aspects. First, Sethi and Somanathan study a group of agents where every agent interacts with everybody else, i.e. interaction is global. In contrast, our model explicitly assumes interaction to be local, an assumption which is motivated by the observation that social mechanisms are most likely to work in local close–knit groups rather than in global anonymous settings. Second, Sethi and Somanathan choose the replicator dynamics determining the evolution of the system. We depart from this evolutionary assumption and, instead, take a closer look on the microlevel of the population. This leads to a more socio–psychologically founded mechanism which is based on education through punishment. Finally, Sethi and Somanathan note that their approach "does not allow for institutional change", where "it is possible for more centralized enforcement mechanisms in the form of explicit laws, policing, and instutionalized punishment to evolve at the local level".[6] On the contrary, our paper can be seen — and in fact should be seen — as a paper on institutional change, since it captures the idea of a (weak) central institution that

[4] Earlier Bankston and Cramer (1974) argued that too severe punishments can cause alienation and thus decrease compliance.

[5] For why local interaction may yield cooperation in dilemma situations when agents are following simple learning rules, see e.g. Eshel, Samuelson, and Shaked (1998).

[6] Sethi and Somanathan (1997), p783.

can make use of a decentrally organized system of social control. In this set–up implications of different law policies can already be addressed explicitly. Moreover, our approach may serve as a starting point for a general discussion of central and decentral mechanisms of control, where the co–existence of social norms and legal rules and the co–functioning of centrally and decentrally organized control is taken for granted.

The paper is organized as follows. In Section 2 we build our model of local control. Section 3 contains the analysis and presents the main results. Section 4 concludes.

2 Model

We consider an infinite population of agents located on the one–dimensional set of integers $\mathbb{Z}$. Identify each agent with his location and denote agents by $x, y, z \in \mathbb{Z}$. Interaction is local. Agents exclusively interact with their two nearest neighbors that are located immediately to the left and to the right of them. Thus, for any agent $x \in \mathbb{Z}$ the set of neighbors is equal to $\{x-1, x+1\}$.

There is a game to be played between any pair of neighbors, which we call the *neighbor game.* The basic ingredient of that game is a situation where individual and collective incentives go into opposite directions. For reasons of simplicity and illustration let us assume that this part of the game is given by a Prisoners' Dilemma game (PD) with a constant gain from defection, where, without loss of generality, payoffs are determined from the payoff matrix given in Figure 1.

	C	D
C	$2, 2$	$0, 3$
D	$3, 0$	$1, 1$

Figure 1: A Prisoners' Dilemma game.

Individual optimization in the PD leads to the inefficient outcome (D, D). Since every agent plays the game with both of his two neighbors he thus receives a total payoff of 2. In order to reach a more efficient outcome via cooperation (C, C) it is clear that somehow individual interests have to be overcome. There exist several possibilities for doing this, all of them employing some mechanism of control. The mechanism we want to analyze in this paper relies on what we call a system of *local control* (LC). In a population with local control there is no institution that takes care of individual cooperative behavior. Instead, agents themselves may control each and punish defective behavior of their neighbors by reporting them to the authorities. Thus, interaction has two stages — the first one, where neighbors just play the game, and the second one, where deviant behavior can be punished.

We assume that agents are one of two types, denoted by A and B. Type A always punishes defective behavior (D) of his neighbors, type B never does. Thus type A is a social controller. Apart from this, both types are the same. In particular, we assume that they are both rational players in the neighbor game which means that both maximize individual payoffs given the types of their neighbors.

We assume that punishment leads to a reduction of the individual payoff to a punished agent by an amount of $p > 0$. Of course, punishment should be costly for the punisher, too. However, as long as we model the punishment decision as a purely norm driven act, where the punisher himself does not take into account his costs of punishing explicitly and as long as the dynamics selecting amongst types does not depend on payoffs, these costs are obviously irrelevant. For the rest of this paper we will stick to this assumption and model punishment as a simple norm driven act.[7]

Given the types of his neighbors an agent chooses between C and D maximizing his individual payoff. Since D is a strictly dominant strategy in the PD, his choice does not depend on the current strategies of his neighbors but only on their types, since types determine whether an agent punishes or not. We assume that the type of an agent can be identified by each of his neighbors and that agents have to fix one action for both encounters, i.e. they cannot treat their neighbors differently. Since defection gives a gain of 1 in each single game with one neighbor, an agent will choose D as long as the value of p times the number of neighbors that are of type A does not exceed his total gain of 2. Thus, if p is small ($p < 1$) it is always optimal for an agent to defect, even if he is punished by both of his neighbors. If p is large ($p > 2$) it is never optimal to defect if at least one neighbor is of type A. In the intermediate case ($1 < p < 2$) the number of A–types determines behavior. If only one neighbor is of type A it is still optimal to defect. If both are of type A it is optimal to cooperate.

This leads us to an important distinction that has to be made when talking about punishment and social control. There exist two different control mechanisms in form of punishment that are involved in our set–up. One form is *executed punishment*, the other *threatened punishment*. To understand this distinction consider an agent who has at least one neighbor of type A. In that case, there is always a threat of punishment. Whether this threat deters defection depends on the amount of punishment, p, and on the question whether also the second neighbor is a controller. If the constellation makes cooperation optimal, the threat implies perfect deterrence and punishment remains only to be threatened. It is not executed. This is what we mean when we speak of threatened punishment (even though punishment starts, of

[7] In a follow–up study we describe a model of social control where agents take into account the costs of punishment explicitly.

course, always as being a threat). In case the constellation does not yield deterrence, the agent, facing the threat of punishment, will defect nevertheless and the punishment will be executed. In that sense, executed punishment and (merely) threatened punishment exclude each other. Executed punishment can only occur after defection, while threatened punishment makes the agent cooperating.

Rather than being exogeneously fixed we assume the LC system to evolve endogenously through time. That is, we want to study decentralized social control where the set of controllers is subject to a permanent change. At any instant in time every agent can become a controller but in the same way, he may also change again and stop being a controller. The channel that governs the switch towards becoming a controller shall be *education.* Precisely, we look at *education through controlling* where the idea we pursue is the following: if a non–controlling agent gets controlled by one of his neighbors, there is a strictly positive probability for him to become a controller, too. At the same time, once an agent has become a social controller there is again a force back towards the situation of not being a controller anymore. This force is going to be modelled as a constant drift. The main object of this paper is to analyze the effect of this simple educational mechanism — first, on the evolution of the LC system itself and second, on the play of the neighbor game.

If we look at the two forms of social control, executed and threatened punishment, it does not seem clear which of these forms may have a greater effect as an educational mechanism. In view of our assumptions on the dynamics of the system of LC we may ask, which of these forms makes an agent under control more likely to become a social controller, too? The answer is ambiguous. For example, one could argue that executed punishment could have the greater educational effect since it involves the actual pain of being punished. On the other hand, one might say that the impact of threatened punishment might be greater as it has the power to alter behavior.

Since it is not obvious which is the correct answer and since, in fact, the correct answer might be case dependent we will consider different situations, allowing an understanding of the complete picture. In particular, we will analyze cases where executed punishment and threatened punishment have similar educational effects, as well as constellations where their educational impacts differ.

Technically, we model the system of local control (LC) as a continuous time Markov process with state space being the collection of all possible arrangements of controllers. Define $X := \{A, B\}^{\mathbb{Z}}$. X is the collection of all possible arrangements of A– and B–types over the population of agents located on the set of integers $\mathbb{Z}$. LC is then defined through a Markov process $\{\sigma_t\}_{t\geq 0}$ where at any time $t \in \mathbb{R}_0^+$, σ_t is an element of X. Denote by $\sigma_t(x)$ the type of agent x given the state of the process σ_t. Since time is a continuous parameter in this model transition probabilities for the Markov process can be described by some real–valued functions $c(x, \sigma_t)$, called flip rates, that depend on $x \in \mathbb{Z}$ and $\sigma_t \in X$. The family of Markov processes we consider in this context belong to the class of interacting particle systems.[8] Flip rates determine the probability that given a state of the process σ_t an agent x 'flips' his type $\sigma_t(x)$ within an infinitesimally short period of time. Precisely,

$$Prob[\sigma_{t+s}(x) \neq \sigma_t(x)] = c(x, \sigma_t) \cdot s + o(s) \qquad \text{for } \; s \downarrow 0, \tag{1}$$

The probability for agent x to flip within a time interval $[t, t+s]$ equals the product of the flip rate at time t times the length of the interval s plus a term vanishing of order s as s goes to zero.

Like in the case of a standard Poisson process, flip *rates* in $[0, \infty]$ and flip *probabilities* in $[0, 1]$ form a monotone relation: the higher the rate the higher is the probability of flipping within a short period of time. For example, a flip rate equal to infinity implies an instantaneous flip. Here the probability of flipping equals one. On the other hand, a flip rate equal to zero corresponds to a no flip situation where the probability of flipping is zero.

As mentioned above we assume the mechanism for an agent to flip from a B–type to an A–type (i.e. from a non–controller to a controller) to be based on education through control by neighbors. The reverse flip from an A–type to a B–type is modelled by a constant drift term. Let $n_t^A(x) \in \{0, 1, 2\}$ denote the number of neighbors of agent x that are of type A at time t. We define flip rates as follows:

$$c(x, \sigma_t) := \left\{ \begin{array}{ccc} f(n_t^A(x), p) & \text{if} & \sigma_t(x) = B \\ \epsilon & \text{if} & \sigma_t(x) = A. \end{array} \right. \tag{2}$$

The parameter $\epsilon > 0$ models the constant drift from an A–type to a B–type. The function f models the flip from a non–controller to a controller, i.e. from a B–type to an A–type. Capturing the educational effect of being controlled by neighbors it depends both on the number of controllers in the neighborhood and on the punishment level p.

In our model we distinguish between executed punishment and threatened punishment by assuming that the educational effect of executed punishment results into a flip rate $\xi > 0$, while the impact of threatened punishment results into a flip rate $\theta > 0$. Since agents are payoff-maximizers in the neighbor game the fact whether punishment is executed or threatened depends exclusively on the number of neighboring A–types and the level of punishment. This leads to the following definition of f:

[8] See Liggett (1985) for an outstanding introduction to this class of models.

$$f(n,p) := \begin{cases} 0 & \text{if} \quad n = 0 \\ \\ \xi & \text{if} \quad p < 1 \wedge n \geq 1 \\ \\ \xi & \text{if} \quad 1 < p < 2 \wedge n = 1 \\ \theta & \text{if} \quad 1 < p < 2 \wedge n = 2 \\ \\ \theta & \text{if} \quad p > 2 \wedge n \geq 1. \end{cases} \tag{3}$$

If there is no A–type among neighbors, ($n = 0$), there is obviously no control and thus no punishment. Therefore flip rates are zero. If there is at least one A–type around, ($n \geq 1$), there is either executed punishment or threatened punishment, depending on the level of punishment and the number of controllers. If the level of punishment is small ($p < 1$) defection is optimal independent on whether n equals one or two. Therefore punishment is executed punishment which leads to a flip rate equal to ξ. If the punishment level lies between one and two, punishment is executed if and only if exactly one neighbor is of type A. In this case the flip rate has again a value of ξ. As soon as both neighbors are controllers ($n = 2$) it is rational to cooperate and then punishment turns into threatened punishment which, by assumption, has an effect of θ. If the punishment level is even larger ($p > 2$) cooperation is optimal if there is at least one A–type around and then punishment is always threatened punishment leading again to a flip rate equal to θ.[9]

We next turn to an analysis of the model.

3 Analysis and Results

The main question is the following: What are the effects of the simple educational mechanisms described above, first on the endogenous evolution of the system of local control (LC) and second on the play of the neighbor game? Obviously, the state of *no–control*, where $\sigma_t(x) = B$ for every agent $x \in \mathbb{Z}$, is an absorbing state. Denote this state by $\emptyset$. We are interested in the probability for LC to get around the absorbing state of no–control. That is we are going to analyze the probability for the set of controllers to be nonempty at any time $t \geq 0$. The next definition clarifies the terms we are going to consider.

Definition 1 *The system of local control (LC)* **survives** *if for every initial state $\sigma_0 \neq \emptyset$ the probability for $\{\sigma_t\}_{t \geq 0}$ not to enter the absorbing state of no–control $\emptyset$ is strictly positive. The*

[9] Note, that we do not consider cumulative punishment effects in this model. If, e.g., $p < 1$ such that every punishment is executed, there is no difference between a situation where only one neighbor is punishing and the one where both are of type A and punish. The only thing that matters, is whether there is *at least one* neighbor that punishes and which kind of punishment is at work.

system **breaks down** *if the state of no–control is reached almost surely for every initial state. LC is* **successful** *if it survives and (local) cooperation can be observed. The system is* **weakly successful** *if local cooperation is observed but survival can not be guaranteed.*

Remark: We will provide conditions for survival and breakdown of LC. Once these conditions are fulfilled it can be shown that there exists a class of initial states for which the probability not to enter the state $\emptyset$ is even equal to one. The class consists of those initial states that almost surely contain infinitely many agents of type A. Examples are Dirac distributions putting probability one on a state with infinitely many agents being of type A or Bernoulli product measures, where every agent independently of other agents is of type A with some strictly positive probability. In view of policy analysis this class appears of particular interest because one can ensure that almost surely the set of controllers is never empty. To guarantee this result, note that no assumption on the local spread, i.e. the density or concentration of A–types within the population has to be made. The sufficient condition on the set of controllers is that it shall have at least some positive measure within the infinite space $\mathbb{Z}$.[10]

We are now going to analyze the possibility for survival and breakdown, success and weak success depending on the values of the parameters in the model, i.e. depending on θ, ξ, ϵ, and p. The analytical result that will work as the main tool is given in the following lemma.

Lemma 2 *For every $\epsilon > 0$,*

(i) there exists a value $b(\epsilon) > 0$ such that the system of local control breaks down if for every $n \in \{0, 1, 2\}$ and $p \in (0, \infty) \setminus \{1, 2\}: \; f(n, p) < b(\epsilon)$,

(ii) there exists a value $s(\epsilon) < \infty$ such that the system survives if for every $n \in \{0, 1, 2\}$ and $p \in (0, \infty) \setminus \{1, 2\}: \; f(n, p) > s(\epsilon)$.

(iii) It holds that $\epsilon \leq b(\epsilon) \leq s(\epsilon) \leq 4\epsilon$.

The proof of Lemma 2 is referred to the appendix at the end of the paper. At this point we would like to mention only that from the proof it can be observed that the statement in Lemma 2 can be strengthened in the sense that both conditions for survival and for breakdown can be relaxed. However, since we are not going to focus on this property in this paper the statement as it is given now is enough.

Lemma 2 puts relative bounds on the rates that govern the flip from a B–type to an A–type in order to observe either a breakdown or a survival of the system of local control. Moreover, these bounds are given in relative terms. Thus, in particular, in order to get survival it is not necessary for the drift term eventually to reach zero. We could state the result the other way round: For every degree of educational effect there exists a strictly positive value such that

[10] Consider any finite translation invariant measure λ on $\mathbb{Z}$. Then for every finite subset $S \subset \mathbb{Z}$, $\lambda(S) = 0$.

survival has a chance if only the drift is smaller than this critical value. This feature seems exceptionally nice since many models in the evolutionary literature focus on results where drift or mutations eventually must decrease to zero. In our model this is not necessary. There are always agents that at some time t refuse to be controllers any longer. Hence, there is a permanent change within the set of controllers. The only thing that matters is that overall there is enough control dispersed among agents in the society.

Lemma 2 suggests that we have to consider four different cases. In the first case educational impacts of both executed and threatened punishment are small, while in the second case both effects are large. In the third case the effect of executed punishment is small but the effect of threatened punishment is large. In the fourth case the effect of executed punishment is large but the impact of threatened punishment is small. We will establish four propositions, each dealing with one of the four cases. This will also bring the punishment level p back into the discussion.

Proposition 3 *If both educational forces as small, i.e. if* $\max\{\xi, \theta\} < b(\epsilon)$*, local control will always break down regardless of the punishment level.*

Proof: The claim follows directly from Lemma 2, part (i). □

Proposition 3 cannot be surprising. If there are no considerable forces turning non–controllers into controllers, given the constant drift towards no–control, controllers will disappear pretty soon. Once they disappeared, it will always be optimal to play defection.

Proposition 4 *If both educational forces are strong, i.e. if* $\min\{\xi, \theta\} > s(\epsilon)$*, local control will survive regardless of the punishment level. However, if the punishment level is too small, i.e. if* $p < 1$*, LC will not be successful. If, instead,* $p > 1$*, LC will always be successful.*

Proof: The first claim follows again directly from Lemma 2, part (ii), while the last two claims follow from the observation that when both neighbors are of type A it is always rational to play defection if $p < 1$ and to cooperate if $p > 1$. □

Again the proposition seems intuitive. If both educational forces are strong more and more agents will become controllers be it through threatened or through executed punishment. Eventually, there will be so many controllers that any serious punishment level ($p > 1$) will ensure non–deviant behavior.

Next we turn to the case where one educational force is strong and the other is weak. Here we will see that it will crucially matter whether in fact executed punishment is the strong force or threatened punishment is the strong force. We start with the latter.

Proposition 5 *If threatened punishment is the strong force while the effect of executed punishment is weak, i.e. if $\theta > s(\epsilon)$ and $\xi < b(\epsilon)$, LC breaks down for $p < 1$, it is weakly successful for $1 < p < 2$, and it is successful for $p > 2$.*

Proof: The claims for $p < 1$ and $p > 2$ are again contained in Lemma 2, part (i) and (ii), since executed punishment is at work if $p < 1$, and threatened punishment applies if $p > 2$, regardless of the number of controlling neighbors. The intermediate case ($1 < p < 2$) is more tricky, because unfortunately Lemma 2 is of no help. In order to see why survival can not be guaranteed consider the situation where a B–type has only one neighboring A–type. Since punishment is executed in this situation and $\xi < b(\epsilon)$, the A–type is more likely to drift to a B–type before education can work leading the B–type to become an A–type. The result is a neighborhood of no–control. On the other hand, breakdown does not happen either. Consider the situation of a B–type surrounded by two A–types. In this case, punishment is threatened punishment and the B–type cooperates. Since by assumption the resulting force is strong, i.e. $\theta > s(\epsilon)$ the B–type has a high probability to become an A–type before one of the A–types drifts into a B–type. Thus, with high probability the neighborhood turns into complete control. Still, if this neighborhood is a local island within a set of B–types, in particular if the minimal distance to the next A–type is greater than two, the situations at the borders of the neighborhood are of the kind described above. In consequence, the island will eventually break down again. This shows that the only forces for survival rely on pairs of A–types that surround a single B–type. This, however, is not enough to guarantee survival. That LC is still weakly successful is based on the fact that whenever both neighbors are of type A it is optimal to cooperate. Thus local cooperation can be observed. □

In order to understand the intuition of the proposition consider a legislator that may influence the value of the punishment level p. Proposition 5 shows that whenever threatened punishment is the strong force it is optimal for the legislator to increase the punishment level. The reason is that both in the neighborgame incentives are moved towards cooperation and the educational impact of punishment is increased which in consequence sustains the system of control. That the latter does not necessarily always hold true can be seen from the following proposition where the asymmetry between ξ and θ is turned the other way round.

Proposition 6 *If executed punishment is the strong force while the effect of threatened punishment is weak, i.e. if $\xi > s(\epsilon)$ and $\theta < b(\epsilon)$, LC survives but is not successful for $p < 1$, it is weakly successful for $1 < p < 2$, and it breaks down for $p > 2$.*

Proof: Similarly to above the extreme claims follow again from Lemma 2, part (i) and (ii). If $p < 1$ agents exclusively control by executed punishment. If $p > 2$ punishment is always threatened punishment. In both cases the number of controlling neighbors is irrelevant. The intermediate case where $1 < p < 2$ is again the complicated one. Consider a neighborhood where a B–type is surrounded by two A–types. In such a situation the B–type cooperates and punishment is threatened, which has an effect of θ. Since θ is small with a high probability one of the A–types will drift into a B–types before the B–type gets educated and flips to an A–type. In consequence, the neighborhood turns into two B–types next to an A–type. But now the B–type in the middle starts defecting and therefore experiences executed punishment by his single neighboring A–type. As executed punishment has a strong force in this set–up it becomes more likely that now the B–type will flip into an A–type before the remaining A–type drifts into a B–type, too. (As his former colleague has done before.) Thus, we get two A–types next to a single B–type. Just as in the case of proposition 5 where it were the pairs of A–types that sustained the spread of controllers it is now a single A–type that builds the stronger force. However, in any situation that involves a pair of A–types surrounding some B–type controllers are extremely vulnerable. Therefore survival in the sense of not reaching the state of no–control with positive probability for every initial state seems out of reach and can not be guaranteed. Still, compared with the situation in proposition 5 chances seem to be better than before. Moreover, by the same reasoning as above cooperation can be observed and thus LC is weakly successful. □

Policy trap

Propostion 6 reveals an important feature of our model that is based on the possibility of type change and the distinction between different forms of punishment as its main characteristic. This feature involves what we call a policy trap.

As before consider again a legislator that may want to change the value of p. Let us suppose that the initial situation is such that $p < 1$. The legislator may see that there is a lot of control in the population, A–types survive since ξ is large enough. However, nobody cooperates since the punishment level is too low. In consequence, he may want to increase p in order to give cooperation a chance. Suppose now that first he increases p to a level where $1 < p < 2$. In this situation cooperation is optimal if and only if both neighbors are controllers. If only one neighbor is of type A defection is still a dominant strategy. Now, if both neighbors are of type A successful control is exercised *(per definitionem)* by threatened punishment. But, by assumption this has a lower educational effect. In fact, these states where two A–types surround one B–type are the most unstable and therefore situations that have high frequency involve only one single controller within a neighborhood. Yet, in these situations it is not optimal to cooperate. Thus, the legislator may come to the conclusion that the punishment level is still too low since most of the situations where control is involved are not successful.

Yet, if $p > 2$ the situation becomes even worse. Although in the short run, the policy looks extremely appealing, since everybody cooperates if only one controller is around, eventually this leads into a trap. If $p > 2$, all control is exercised by threatened punishment. By assumption this has an educational effect only of θ, which is not enough to have a B–type flipping to an A–type with a higher probability than his controllers drifting to a B–type. Still, this is necessary to give the system of LC a chance for survival, at all. In consequence, in the long run the system breaks down. All A–types disappear, i.e. nobody is left to control. Hence, nobody cooperates. The situation is worse than before and even worse than the one we started from. While at the beginning there was at least a system of control, which, however, was not successful, we are now in a state where the system is not only not successful but where it does not even exist anymore.

This shows that as long as the legislator has no influence on the educational forces themselves the best he can do is keeping the punishment level between 1 and 2. Although he pays for obtaining local cooperation by loosing survival of LC he nevertheless stays away from a complete breakdown of the system. In this set–up the second–best solution in form of weak success of LC is all he can get.

Figure 2 summarizes our findings.

	$p < 1$	$1 < p < 2$	$2 < p$
(C1)	−	−	−
(C2)	◦	++	++
(C3)	−	+	++
(C4)	◦	+	−

(− = breakdown; ◦ = survival; ++ = success; + = weak success;
C1: $\max\{\xi, \theta\} < b(\epsilon)$; C2: $\min\{\xi, \theta\} > s(\epsilon)$; C3: $\xi < b(\epsilon) \leq s(\epsilon) < \theta$; C4: $\theta < b(\epsilon) \leq s(\epsilon) < \xi$.)

Figure 2: The effects of the punishment level p in the four different cases.

4 Conclusion

We have analyzed a system of so–called local control (LC), where the task of enforcement of punishment lies in the hands of individuals in the society rather than in the hand of a centrally organized institution. Having in mind the idea of social engineering we have studied a world where the survival of LC is made as difficult as possible. First, all agents are utility maximizers and second, there is a constant (psychological) force producing a drift towards a state of no–control. In spite of these assumptions it is possible to give conditions such that LC has, indeed, a strictly positive chance of survival for every initial state of the society. The important interactive mechanism between agents that produces this result is education through punishment. Non–controlling agents may become controllers when they experience control themselves. With this respect, we have distinguished between executed and threatened punishment. Although survival of LC is possible, the alternative, breakdown of LC, must not be ignored. For every degree of the educational effect of experienced control there exists a sufficiently strong drift towards no–control that eventually leads the system to a complete breakdown of control. This feature becomes particularly important in a situation where executed punishments have stronger impacts than mere threats. In this case the interplay between types and incentives form a policy trap, where any legal or social policy focussing on punishment levels alone can do more harm to the situation than first economic insights would have let expected.

Appendix

Proof of Lemma 2: The system of LC $\{\sigma_t\}_{t\geq 0}$ with flip rates defined via function f (see equation (2)) is a nearest–particle system with parameters

$$\beta(l,r) = \begin{cases} \frac{f(2,p)}{\epsilon} & \text{if} \quad l = r = 1 \\ \frac{f(1,p)}{\epsilon} & \text{if} \quad l = 1 \ \text{ or } \ r = 1 \ \text{ and } \ l + r > 2 \\ 0 & \text{if} \quad l \geq 2 \ \text{ and } \ r \geq 2. \end{cases} \tag{4}$$

The value of $\beta(l,r)$ equals, up to multiplication by ϵ, the value of the flip rate $c(x,\sigma_t)$, when $\sigma_t(x) = B$, l is the distance to the nearest A–type to the left and r is the distance to the nearest A–type to the right.[11]

For $1 \leq n < \infty$ consider the collection of numbers

$$b(n) = \sum_{l+r=n+1} \beta(l,r). \tag{5}$$

Obviously,

$$b(n) \leq 1 \ \text{ for all } \ 1 \leq n < \infty \tag{6}$$

$$\text{iff} \ \ \frac{f(2,p)}{\epsilon} \leq 1 \ \text{ and } \ \frac{2f(1,p)}{\epsilon} \leq 1. \tag{7}$$

Theorem 5.5., chapter VII of Liggett (1985) assures that the system breaks down if (8) holds. Thus equivalently, condition (9) is sufficient to ensure a breakdown of the system of LC. This proves already the existence of a value $b(\epsilon) > 0$ such that the system breaks down if $f(n,p) < b(\epsilon)$ for every $n \in \{1,2\}$. So far, we know from (9) that $b(\epsilon) \geq \frac{\epsilon}{2}$. However, it is possible to improve this bound. The idea is to compare the LC with some other nearest–particle system for which the breakdown condition is already better known. In fact, this approach will allow us to prove the second claim, too. The other particle system is the contact process. (See Liggett (1985), chapter VI.)

The contact process is a nearest–particle system with parameters

$$\tilde{\beta}(l,r) = \begin{cases} 2\lambda & \text{if} \quad l = r = 1 \\ \lambda & \text{if} \quad l = 1 \ \text{ or } \ r = 1 \ \text{ and } \ l + r > 2 \\ 0 & \text{if} \quad l \geq 2 \ \text{ and } \ r \geq 2. \end{cases} \tag{8}$$

[11] A nearest–particle system is commonly defined with a constant drift term equal to 1 instead of ϵ. The multiplication of flip rates by a constant term — in our case by ϵ — has no qualitative effect other than a change of the time scale.

Now the better bound for $b(\epsilon)$ relies on the fact that the system of LC is dominated by the contact process with parameter λ if both $\frac{f(1,p)}{\epsilon} \leq \lambda$ and $\frac{f(2,p)}{\epsilon} \leq 2\lambda$. Domination means that for any arrangement of types, every B–type is at most as likely to flip under the LC system as under the contact process. That is, $\beta(l,r) \leq \tilde{\beta}(l,r)$ for any selection of l and r.

For the contact process it is well–know that there exists a critical value λ^* such that for every $\lambda < \lambda^*$ the process breaks down and for every $\lambda > \lambda^*$ the process can survive. In the one–dimensional case it is possible to show that this critical value has a lower bound of 1.18 and an upper bound of 2.[12] Moreover, if $\lambda > \lambda^*$ any translation invariant initial distribution that puts mass zero on the no–control state $\emptyset$ converges weakly to an invariant distribution ν_λ that has the property that with strictly positive probability any agent is of type A and, in consequence, with probability one at least one agent in the population is of type A. (See chapter VI of Liggett (1985).)

Thus taking 1 as a lower bound for λ^*, $f(1,p) < \epsilon$ and $f(2,p) < 2\epsilon$ is a sufficient condition for the LC system to be dominated by a contact process that breaks down. This proves that $b(\epsilon) \geq \epsilon$.

The second claim follows from the same dominance argumentation, this time the other way round. We take 2 as an upper bound for λ^*. Then $f(1,p) > 2\epsilon$ and $f(2,p) > 4\epsilon$ are sufficient conditions for the LC system to dominate itself a contact process that can survive. Hence, the LC system can survive, too, which proves already the existence of a value $s(\epsilon) < \infty$ such that the system can survive if $f(n,p) > s(\epsilon)$ for every $n \in \{1,2\}$. Furthermore, it follows also that $s(\epsilon) \leq 4\epsilon$. □

[12] An approximation done by Brower *et al.* (1978) obtains a value of $\lambda^* \approx 1.6494$.

References

Akerlof, G.A. and Dickens, W.T. (1982) "The Economic Consequences of Cognitive Dissonance", *American Economic Review*, 72(3), 307-319.

Bankston, W.B. and Cramer, J.A. (1974) "Toward a Macro–Sociological Interpretation of General Deterrence", *Criminology*, 12(3), 251-280.

Beyleveld, D. (1980) *A Bibliography on General Deterrence Research*, Saxan House, Westmead.

Brower, R.C., Furman, M.A., and Moshe, M. (1978) "Critical exponents for the reggeon quantum spin model", *Physics Letters B*, 76, 213-219.

Durkheim, É. (1893) *De la Division du Travail Social*, Alcan, Paris. (Translated by Halls, W.D., *The Division of Labor in Society*, 1984, The Free Press, New York.)

Ellickson, R.C. (1991) *Order without Law*, Harvard University Press, Cambridge, MA.

Eshel, I., Samuelson, L. and Shaked, A. (1998) "Altruists, Egoists, and Hooligans in a Local Interaction Model", *American Economic Review*, 88(1), 157-179.

Hay, J.R. and Shleifer, A. (1998) "Private Enforcement of Public Laws: A Theory of Legal Reform", *American Economic Review*, 88(2), 398-403.

Humboldt, W. von (1792) "Ideen zu einem Versuch, die Grenzen der Wirksamkeit des Staates zu bestimmen", in Humboldt, W. von, *Schriften zur Antropologie und Geschichte*, 1980, Cotta, Stuttgart, 56-233. (Translated by Burrow, J.W. (Ed.), *The Limits of State Action*, 1993, Liberty Fund, Indianapolis.)

Kahan, D.M. (1996) "What Do Alternative Sanctions Mean?", *University of Chicago Law Review*, 63(2), 591-653.

Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. (1982) "Rational Cooperation in the Finitely–Repeated Prisoners' Dilemma", *Journal of Economic Theory*, 27, 245-252.

Liggett, T.M. (1985) *Interacting Particle Systems*, Springer-Verlag, New York.

McMillan, J. (1994) "Selling Spectrum Rights", *Journal of Economic Perspectives*, 8(3), 145-162.

Okada, A. (1997) “The Organization of Social Cooperation: A Noncooperative Approach”, in Albers, W. et al. (Eds.), *Understanding Strategic Interaction, Essays in Honor of Reinhard Selten*, Springer-Verlag, 228-242.

Roth, A.E. (1990) “New Physicians: A Natural Experiment in Market Organization”, *Science*, 250, 1524-1528.

Roth, A.E. and Peranson, E. (1997) “The Effects of the Change in the NRMP Matching Algorithm”, *Journal of the American Medical Association*, September 3, 729-732.

Salem, R.G. and Bowers, W.J. (1970) “Severity of Formal Sanctions as a Deterrent to Deviant Behavior”, *Law and Society Review*, 5(1), 21-40.

Sethi, R. and Somanathan, E. (1996) “The Evolution of Social Norms in Common Property Resource Use”, *American Economic Review*, 86(4), 766-788.