Journal of Economic Interaction and Coordination
https://doi.org/10.1007/s11403-021-00316-7

REGULAR ARTICLE

The evolution of norms within a society of captives

Chad W. Seagren[1] · **David Skarbek**[2]

Received: 27 April 2020 / Accepted: 16 January 2021

Abstract
How do norms evolve when people have no choice to opt out of social interactions? One example of such a setting is prison. Past research usually relies on ethnographic work to understand the emergence and maintenance of norms among prisoners. We instead use this rich qualitative literature to inform an agent-based model to demonstrate how norms evolve in response to demographic changes in prison. In the model, agents play a one-shot, though possibly repeated, prisoner's dilemma with other agents. Agents lack the ability to decline to play with their selected opponent. We consider tag-mediated play and norm enforcement as mechanisms to facilitate prisoner cooperation and to examine the effects of increasing prison populations and increasing ethnic heterogeneity on the maintenance of cooperative norms. We also calibrate the model with empirical data from the California prison system. Parameters of the model correspond to demographic changes between 1951 and 2016, where the size of the prison population increased 14-fold and ethnic heterogeneity by 30%. Simulation results show that such changes dramatically decrease levels of cooperation and compliance. These results are consistent with the actual observed breakdown of the cooperative norms in California prisons.

Keywords Self-governance · Norms · Prison · Agent-based model

JEL Classification K4 · P48 · P37

✉ Chad W. Seagren
cwseagre@nps.edu

David Skarbek
Davidskarbek@gmail.com

[1] Graduate School of Defense Management, Naval Postgraduate School, Monterey, CA, USA

[2] Department of Political Science and the Political Theory Project, Brown University, Providence, RI, USA

Published online: 09 April 2021

1 Modeling self-governance

Social scientists have long studied how cooperation can emerge among people in the absence of effective and credible third-party enforcement. Social dilemmas often lead to situations where mutually beneficial outcomes fail to materialize. While reliance on a third party, such as the state, is one possible solution to social dilemmas, many others exist as well. This issue has been studied in numerous ways, including both theoretical (Dixit 2004) and experimental approaches (e.g., Powell and Wilson 2008). In addition, there is also now a large empirical literature studying how people engage in self-enforcing exchange in different historical contexts (Ellickson 1991; Greif 1993; Ostrom 1990; Leeson 2014; Stringham 2015; Skarbek 2020).

Two of the key mechanisms found in these studies are: (1) the ability to choose with whom to interact and (2) the ability to exclude non-cooperative individuals. One limitation of these important works is that they tend to focus on cases where cooperation is more likely to emerge: among wealthy, high-status individuals in business. Likewise, in past computational and experimental studies, the ability to exit facilitates cooperative interactions with others (Schuessler 1989; Orbell et al. 1984; Vanberg and Congleton 1992). Instead, we follow Leeson (2007a, b, 2009, 2010, 2014) who emphasizes the study of cases where cooperation might be undermined by violent actors. We contribute to the literature on self-governance by modeling the interaction among people who do *not* have the ability to opt out: prisoners.

A common finding in classic, scholarly studies of prison life is that norms often emerge spontaneously within a prisoner community to provide order (Radford 1945). Prisoners often refer to these informal social rules as the "convict code." Seminal studies in sociology and criminology describe these norms and identify how they serve as a source of social control within the society of captives (Clemmer 1940; Sykes 1958). Because prisoners cannot simply leave the prison, the exit option cannot facilitate cooperation—unlike in many existing studies of self-governance (Tullock 1985; Stringham 2015). We use our agent-based model to study the robustness of norms to facilitate cooperation among prisoners. This contributes, more generally, to the large body of work that examines the robustness of informal institutions in governing social and economic affairs (Powell and Stringham 2009).

While the issue of exit costs and norm formation applies in many settings (for example, within the military, aboard ships at sea, restrictions on emigration from conflict regions, etc.), we have characterized the model specifically as a representation of prisoner interactions for three reasons. First, it allows us to contribute to the large literature on prison social order that exists in anthropology, criminology, and sociology. Classic works rely on ethnographic and qualitative research methods (Clemmer 1940; Sykes 1958; Irwin 1980; Irwin and Cressey 1962). These landmark studies tell us much about prison life, but they are subject to confounding by numerous variables and are not focused on identifying causal mechanisms (as in analytical narratives, such as Greif 2006). Second, an agent-based model simplifies and formalizes many facets of the real world to enhance our understanding of key issues. Our work complements these past approaches. Importantly, it is not feasible to vary

prison populations exogenously for the purpose of testing these relationships.[1] An agent-based model, however, allows us to better understand the relationships identified in this early work. Second, the rich ethnographic literature provides support for the accuracy of the *critical* assumptions of our model (Rodrik 2015, 25–29). Finally, we can informally assess the plausibility of the model by comparing it to historical findings on prisons. In particular, we combine data on the California prison system with our model to simulate changes in norm following observed historically. While the data do not exist to test the model with standard statistical approaches, the available evidence is sufficient to show that it is consistent with the model. This finding is important because it links increases in the size of the prison population with the breakdown of norm following among prisoners. This adds an additional element of concern to the problematic, current state of American mass incarceration.

We build on past studies of agent-based models of repeated Prisoner's Dilemma (PD) games by leveraging tag-mediated play along with a norm-enforcement algorithm.[2] Holland (1993) first developed the concept of tags, which influenced Riolo (1997), Riolo et al. (2001), and Hales (2001). In these models, agents take cues from the information presented in the players' tags to decide which strategy to employ, or even to play at all. The cognitive algorithm we employ is similar to that described in Axelrod (1986). Essentially, the agents in our model rely primarily on tags to decide how to interact with strangers, and they are afforded the opportunity to punish non-compliance that they observe between other agents. This enables them to generate cooperative behavior in one-shot, though possibly repeated, games without the luxury of avoiding or refusing to play any opponents, such as in Janssen (2008).[3]

The notion that individuals' respective group affiliations might affect their willingness to cooperate with each other relates to research in the economics of identity. Akerlof and Kranton (2000, 2005) are seminal works in this area, as well as Goette et al. (2006), Benjamin et al. (2010), and Chen and Chen (2011). A related literature exists in psychology based on social identity theory developed by Tajfel (1974) and Tajfel and Turner (1979). See De Cremer and Van Vugt (2002) and Fischer (2009) for recent relevant examples. Finally, Fehr and Gächter (2000) has inspired a growing literature on norm enforcement in the behavioral economics literature. Examples in this stream include Cubitt et al. (2011) as well as Carpenter and Matthews (2009).

We develop an agent-based model in which prisoners interact in a classic one-shot Prisoner's Dilemma game. Agents in the model possess tags, which might be considered outward manifestations of an agent's reputation for adhering to the convict code, as well as other easily apprehensible physical features such as ethnic

[1] It is notoriously difficult to perform any human subject experimentation in this area. See, for example, the Stanford Prison Experiment, described in Haney et al. (1972).

[2] De Marchi and Page (2014) provides an excellent recent survey of the use of agent-based models in studying political and social questions.

[3] This paper also contributes to the literature that applies modeling and simulation to criminal justice issues. A seminal paper in this literature is Joshua Epstein's (2002) model of civil violence. Goh et al. (2006) and Zou et al. (2012) provide refinements to that model. See also Melleson et al. (2012) on burglary, Austin et al. (2012) on street gang affiliation, and Tako and Robinson (2010) on modeling the U.K. prison population.

group. Agents decide upon their PD strategies in part based on their opponent's tags. In contrast to most models that rely on tag-mediated strategies, the agents in this model do not decide with whom to associate and cannot refuse to participate in the game once their opponent is (randomly) selected. Despite this obstacle to cooperative behavior, we find a substantial domain of simulation parameters in which cooperation among prisoners is widespread. However, we show that cooperative behavior is vulnerable to changes in size and composition of the population. Finally, we use data from the California Department of Corrections and Rehabilitation to calibrate the model over a long historical timeline. Between 1951 and 2016, the size of the California prison population increased 14-fold and ethnic heterogeneity by 30%. Our simulation results are consistent with the actual observed breakdown of the convict code in California prison, which helps us to validate the model and to demonstrate the robustness of our findings.

2 The convict code: prisoner norms

One of the most prominent themes in the literature of prison social order is the existence and role of the convict code. It is one of the earliest areas of focus within the literature (Clemmer 1940; Sykes 1958). A large body of work documents its existence, describes its content, and debates its permanence (Mitchell et al. 2016). These studies are overwhelmingly based on qualitative evidence, including ethnography, interviews, surveys, and participant observation.

The convict code consists of a system of informal norms that prisoners are expected to adhere to and to enforce. The norms govern economic and social interactions within prisoner society (Sykes and Messinger 1962; Williams and Fish 1974). It provides a source of governance when prison officials either cannot or will not govern effectively.[4] It is not a written document and the content and emphasis varies to some degree across prisons; however, there are several key components that are consistent across nearly all prisons studied.

In a classic article, Sykes and Messinger (1962: pp 401–403) describe these prisoner norms. First, the code admonishes prisoners to "Never rat on a con" and "Don't interfere with inmates' interests" (Sykes and Messinger 1962: 402). Second, the code encourages prisoners to resolve disputes with fellow prisoners without excessive emotion. For example, when disputes arise, the appropriate norm is "Don't lose your head" (Sykes and Messinger 1962: 403). The code discourages taking advantage of other prisoners, with norms that tell prisoners "Don't exploit inmates" (Sykes and Messinger 1962: 403) and to share good fortune with others in reciprocal ways. Fourth, the code encourages prisoners to maintain their strength in the face of hardship: "don't be weak." Finally, the norms prohibit cooperating in any way with

[4] Leeson (2007a, b) and Murtazashvili and Murtazashvili (2015) emphasize that it is often too costly to rely on government to enforce rights. There is also a large literature examining the conditions under which anarchy delivers desirable social and economic outcomes (Powell and Stringham 2009; Mildenberger 2015; Luther 2015).

the prison officials. The meta-norm of the convict code is that prisoners should support other prisoners in nearly all cases, never aid officials, and punish prisoners who defy the code.

Prisoners who abide by the code gain the respect of their fellow prisoners. Being in good social standing means that a prisoner has the mutual support of his peers. Prisoners who fail to uphold the code often experience negative attention from other prisoners, which ranges from verbal reprimand to extreme physical violence (Bowker 1980). Prisoners who regularly violate the code find themselves ostracized and subject to victimization. The code enables prisoners to establish order in their society, improve the security of their persons and property, and to capture gains from trade.

Prisoner culture has a jargon for describing prisoners and the extent to which one complies with the code. For example, a "Right guy" is a highly respected prisoner who is widely known to behave in accordance with the convict code (Williams and Fish 1974). Alternatively, an "Outlaw" or "Rat" is a prisoner who is not respected and is known to regularly violate the convict code. Finally, "Con politicians" are known to adhere to the code only intermittently and when it suits their self-interest. In our model, we examine the extent to which prisoners employ the "Right Guy" strategy.

The folk theorem suggests that a possible equilibrium in indefinitely repeated interactions is mutual, reciprocal cooperation (Fudenberg and Maskin 1986; Tullock 1985). However, folk theorem solutions are less likely to arise in prison for three reasons. First, it is often common knowledge when a prisoner will be released, establishing a finite end to interactions. Second, prisoners cannot credibly employ trigger strategies—such as threats to refuse all social or economic interactions. Prison life is defined by its involuntary association (Rudoff 1964). Finally, future benefits are most alluring to patient people, and prisoners tend to have higher discount rates than the general population (Avio 1998; DiIulio 1996; Pratt and Cullen 2000).

Historically, in California, the convict code was the primary governance structure until the late 1950s and 1960s (Irwin 1980). After that, studies find that the effectiveness of the code to exert social control declined substantially (Hunt et al. 1993). This change has been attributed to two main factors (Skarbek 2012, 2014, 2016). First, there was a dramatic increase in the size of the prison population. This made it more difficult to keep track of other prisoners' reputations, and it increased the likelihood of free riding on enforcement of the code. Second, there was an increase in the ethnic heterogeneity of the prisoner population, which undermined social cooperation. Our model examines the effects of these demographic changes in greater detail, and our findings are consistent with the notion that factors such as population size and composition eroded and ultimately led to the extinction of the convict code.

3 An agent-based model of prison life

We use the RepastJ libraries to implement our agent-based model in Java (North et al. 2013). The agents reside on a two-dimensional grid and move to a vacant, randomly selected cell adjacent to their current position at the beginning of each

Table 1 Prisoner instance variables

Parameter	Details	Explanation	Type
Location	(x, y) coordinates	Agent parameters necessary for maintaining position in environment	Movement parameter
Heading	cardinal direction		
TagString	String of binary digits	tagString identifies agent to others	Interaction parameter
Utils	integer	Measure of success in game-play	Evolutionary algorithm
Strategy array	Array of binary digits	Four digit array that contains agent's strategy	

time-step. Next, agents scan their Moore neighborhood and play a one-shot PD with any and all neighbors. Each game an individual agent plays is independent of the others, so an agent who faces multiple opponents during a given time-step may play different strategies (cooperate or defect) against each. While agents have no memories of past interactions, they do possess tag strings that other agents observe and evaluate for similarity. The agent selects his strategy on the basis of his own, possibly unique, strategy array.

Agents who are not engaged in a PD but witness the play of other agents may take the opportunity to punish an agent that the witness believes acted improperly. Further, agents may act as meta-witnesses who observe the behavior of witnesses. Meta-witnesses may opt to punish witnesses who fail to punish agents for improper behavior. This process is similar to Axelrod's (1986) norm-enforcement model.[5] In our model, the agents deterministically employ the relevant element of their strategy array, rather than probabilistically as in Axelrod. The simulation then advances to the next time-step. Agents in our model are eventually allowed to evolve their strategy choices, which enables widespread norm enforcement and meta-norm enforcement to emerge spontaneously (or not) under various conditions.

We imbue agents with a number of instance variables. In addition to those necessary to place the agent in a location in space and enable it to move around, parameters that govern their interactions with other agents and those involved in the evolutionary algorithm are also necessary. We outline the most important instance variables in Table 1.

The *tagString* parameter is a string of bits that other agents observe and take into consideration when deciding how to interact with the agent. In the analysis below, agents observe their opponent's entire tag string and interpret it without error, though the model is capable of a number of different settings in this regard. Agents deem each other sufficiently similar if the proportion of bits they share in common is above a particular threshold. For example, Fig. 1 contains the tag strings for two

[5] See Prietula and Conway (2009), Kendal et al. (2006), and Mahmoud et al. (2012) for other examples of Axelrod-inspired models of meta-norm emergence. See Horne (2001), Bendor and Mookherjee (1990), Sampson et al. (1997), Carpenter and Matthews (2010), and Fehr and Fishbacher (2004) for discussion of theoretical and experimental examples of third-party norm enforcement. See Kusakawa et al. (2012) for an example of an experiment in which the presence of a witness helps to encourage cooperative behavior in a one-shot PD.

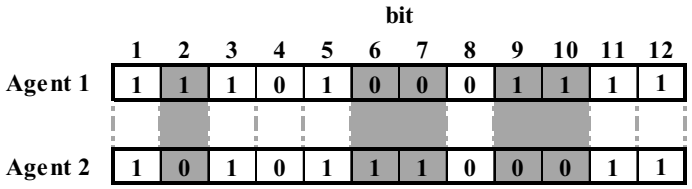

Fig. 1 Tag string example

Table 2 Agent strategy types

Type	PD similar	PD dissimilar	Norm enforce	Meta-norm enforce
0	C	C	E	E
1	C	C	E	N
2	C	C	N	N
3	C	D	E	E
4	C	D	E	N
5	C	D	N	N
6	D	C	N	N
7	D	D	N	N

notional agents. The agents share seven bits (in white) and therefore have a similarity of 0.583. Thus, if the model parameter *similarity_threshold* was set to, say 0.8, these agents would not consider each other similar. But if *similarity_threshold* were set to 0.5, then they would be deemed similar.

Another critical individual parameter is the strategy array the agent employs. The eight different strategy types are shown in Table 2. When the agent plays a PD with another agent, and the agents have sufficiently similar tag strings, the agent cooperates or defects according to the first element of its strategy array. The second element of the strategy array governs play with dissimilar agents. The parts of the strategy that dictate agent behavior in their capacities as a witness or meta-witness are found in the third and fourth elements. Strategies that enforce the norm as witnesses have an *E* in the norm enforce column, while those that elect not to enforce the norm in such cases have an *N*. Similarly, the column meta-norm enforce indicates whether the agent enforces the norm as a meta-witness by punishing witnesses who fail to uphold the norm. We make the assumption that agents are not hypocritical in their decisions to enforce a norm they are unwilling to adhere to themselves, and therefore do not allow a strategy such as (D, D, E, E). One justification for this is that a rat may lack sufficient social standing to attempt to effectively punish another inmate for an infraction.

In our model, the PD serves as a proxy for a typical interaction a prisoner might have with another prisoner. It is helpful to think of the PD in this application as a voluntary exchange of contraband. A prisoner cooperates with others when, for example, he sells an accurate quantity and quality of drugs or when he buys such

Fig. 2 Normal form prisoner's dilemma with payoffs

P1, P2		P2	
		Cooperate	Defect
P1	Cooperate	R, R	S, T
	Defect	T, S	P, P

drugs on credit and follows through to repay his debts. A prisoner defects when he enters into the agreement through fraud or deception, or when he employs violence to seize the product or payment. In addition to an exchange of contraband, the PD could represent any interaction that presents a prisoner with an opportunity to take advantage of a fellow prisoner. Examples of such opportunities may include a prisoner that has information they could report to the guards regarding a fellow prisoner who broke the rules, or a prisoner that notices an opportunity to steal a fellow prisoner's property. We do not necessarily wish to impart a normative judgment on the appropriateness of these transactions from a legal or regulatory point of view. Indeed, many of the transactions we envision may be illegal on the outside, as well. A cooperative outcome, however, does have important implications for the inmates' personal security and property rights. Society at large certainly has an interest in ensuring that prisons are orderly institutions. If inmates can provide a significant proportion of this security on their own, they reduce the burden placed on the guards for providing it.

Figure 2 outlines the payoffs to the PD in the model. The magnitudes of the payoffs are in the typical descending order: temptation (T) > reward (R) > punishment (P) > sucker (S). This order ensures that defect is a strictly dominant strategy for each player. We systematically vary the magnitude of these parameters in our experiments, but maintain this order.

When played in the single-shot setting, the classic PD results in a general unwillingness among the players to cooperate. In fact, in many models, the ability of agents to decline to play other agents is a mechanism through which cooperative behavior emerges (Tullock 1985). However, as discussed above, the inability to select those with whom one interacts is a central feature of prisoner life.

A population of prisoners that exhibits widespread cooperative behavior in this context is consistent with the convict code. The convict code holds that an individual prisoner should, in general, cooperate with other prisoners. Sykes and Messinger (1962: 402–403) note that the convict code contains several directives such as "Don't break your word; don't steal from the cons; don't sell favors; don't be a racketeer; don't welsh on debts," as well as "don't interfere with prisoners'

interests" and "never rat on a con." To defect in the PD violates one or more of the tenants of the convict code.

Our rendering of the convict code is that individual agents should cooperate with all agents regardless of degree of similarity and, as witnesses, they should discipline any agents they observe who fail to live up to the code. For our purposes, agent Type 0 (from Table 2, page 10) embodies these behaviors and might be termed the "Right guy" strategy (Sykes and Messinger 1962: p 404). Agents who employ strategy Type 3 embody these behaviors, with the minor exception they defect against dissimilar agents. Strategies 6 and 7 embody the behavior of "rats" or "outlaws" to varying degrees (Sykes and Messinger 1962: p 405). For purposes of measuring compliance, we consider strategy types 0, 1, and 2 as generally compliant with the code, due to the fact that they cooperate with both similar and dissimilar agents in the PD. In addition to those three strategy types, we consider types 3, 4, and 5 to be generally cooperative, because all six *at least* cooperate with similar agents.

Two mechanisms primarily enable the emergence of convict code compliance and cooperation in our artificial society. First, agents' tag strings provide other prisoners with relevant information, such as group affiliation. An agent may use the measure of similarity shared with its opponent as a means for deciding how to treat him, but keep in mind that in all cases the agents must complete play with their opponent.

The second mechanism is the ability for witnesses to such games to punish those who fail to abide by the convict code. During each time-step, agents not currently involved in playing a game with another agent could serve as witnesses, if they are in close proximity to the game. Agents may be in sufficiently close proximity to witness more than one game, but the one game they witness for purposes of norm enforcement is randomly selected. If the witness observes that one or more players of the PD defects, they will discipline the code violators. When an agent is disciplined in such a manner, they suffer damages equal to D utils. The witness who inflicted the discipline suffers a cost of E utils. In addition, agents not otherwise involved in a game or as a witness are candidates to become meta-witnesses. A meta-witness observes whether a *witness* elects to punish non-compliance in the PD. A witness who neglects to punish non-compliance is himself non-compliant with the code. An agent may be a meta-witness to exactly one other agent per time-step. As above, the subject of the discipline loses D utils, while the disciplining agent loses E.

As Axelrod (1986) finds in his model, we observe the emergence of relatively robust cooperation and behavior consistent with the convict code over a relatively wide range of parameter values and find the role of witnesses and meta-witnesses to be critical in this formation. While willingness to enforce the norm, or the meta-norm, is hard-wired in the sense that elements for both are found in a set of possible strategies, agents select their strategies through an evolutionary algorithm and are entirely free to select strategies that avoid the norm-enforcement role or discard them if the agent finds them to be disadvantageous.

Figure 3 outlines the events executed in each time-step. Every agent moves and scans their neighborhood in search of players, but subsequent events are contingent on the existence of neighbors or witnesses. And while witnesses are identified

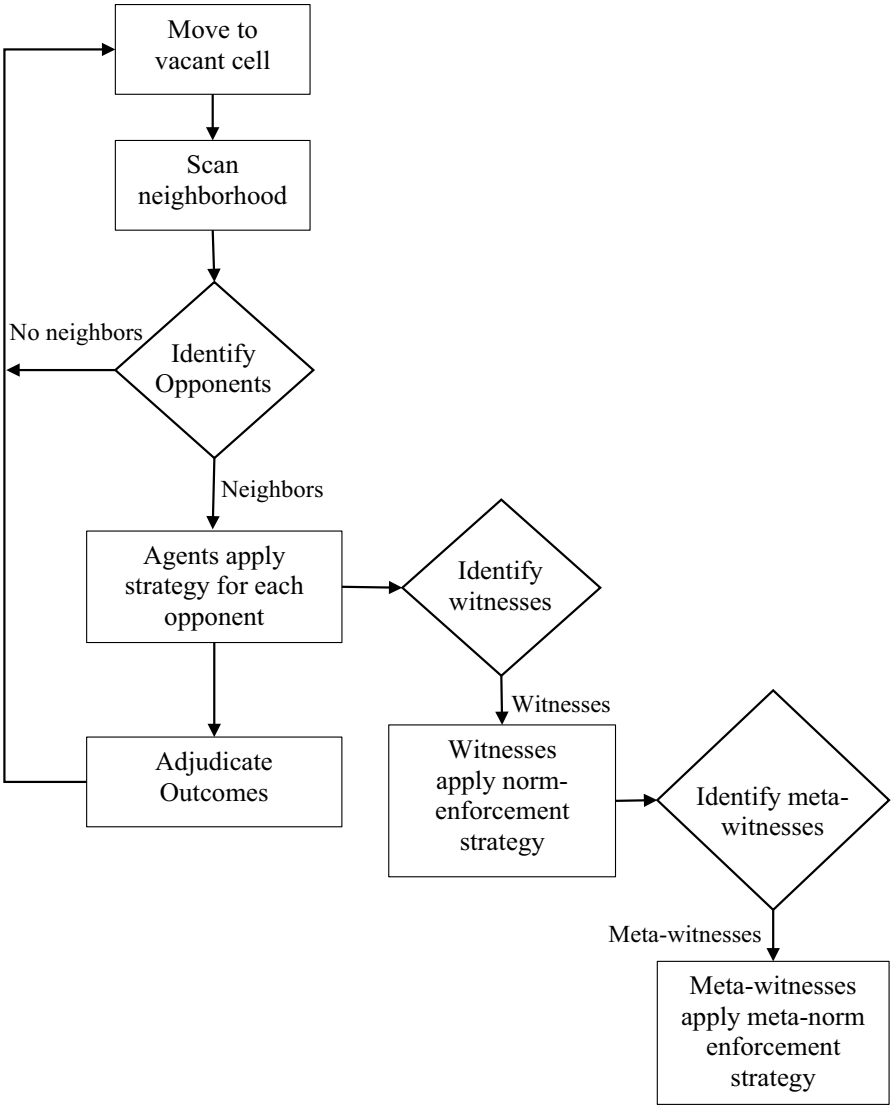

Fig. 3 Event schedule for each time-step

algorithmically through the players, players do not take the existence of witnesses into account directly when selecting their strategies.

At the end of a certain number of time-steps, which we call a generation, we allow each agent the opportunity to change his strategy using a simple hill-climbing algorithm. Upon initialization, agents are randomly assigned a strategy array which becomes their incumbent strategy. As long as the incumbent strategy demonstrates performance (in terms of total utils) that is as good or better relative to their experience in the previous generation, the strategy remains the incumbent. If an incumbent

Table 3 Parameter settings for initial case

Parameter	Description	Value
Generation duration	Time-steps per generation	500
Number of Inmates	Number of agents	365
Size of tagString	Length of bits in tagString	28
Payoff T	Temptation payoff	7
Payoff R	Reward payoff	5
Payoff P	Punishment payoff	3
Payoff S	Sucker's payoff	0
Payoff D	Penalty when witness punishes defection	−7
Payoff E	Cost of enforcement to witness	−2
Discipline	Witnesses punish defection	On
Meta-Discipline	Witnesses punish witnesses who fail to punish	On
Witness range	Range for witnesses observing PD games	5
Meta-witness range	Range for witnesses observing other witnesses	3
Evolve Rule	Agents' strategies allowed to evolve	On
Evolve tolerance	Amount of improvement that candidate strategy must exhibit	1.14

fails to do better, a candidate strategy in the neighborhood of their current strategy is tested out.[6] If the candidate strategy does better than the incumbent in the next generation, it is adopted as the new incumbent. If not, the agent reverts back to the incumbent strategy for another generation.

4 A demonstration of a collapse of the convict code

Before examining the output from our several experiments and calibration effort, it is helpful to discuss a single case in which high levels of cooperation are achieved in an artificial society and then systematically undermined by changing the characteristics of the population. In this section, we walk through a notional case in order to introduce the reader to more of the model parameters, the response variables we measure, the treatments we employ, and the hypotheses we examine. We keep the following parameters in Table 3 constant for the below scenario:

We select this particular set of parameters entirely due to the fact they enable our artificial society to spontaneously achieve cooperative behavior consistent with the convict code.

Our primary measures of effectiveness in this section involve the number of agents whose current strategy types abide by the convict code to some degree. We

[6] We define "doing better" as #utils current generation > #utils previous generation * tolerance. The greater the tolerance factor, the more certain the agent is that the new strategy is better. The tolerance factor helps to encourage stability on the margin, so the agent doesn't cycle endlessly over a set of nearly optimal strategies.

define an agent as cooperative if the agent's strategy involves cooperating with similar agents (any agent with strategy type 0, 1, 2, 3, 4, or 5). We define an agent as compliant with the convict code if they cooperate with both similar and dissimilar agents (i.e., types 0, 1, or 2).

Thus, the proportion of cooperative agents (p) is given by:

$$p = \frac{\#of\ coopertive\ agents}{numInmates}$$

The proportion of compliant agents (c) is given by:

$$c = \frac{\#of\ compliant\ agents}{numInmates}$$

We use these measures both directly and as a means to classify particular aggregate outcomes. For example, we classify the state of the artificial society at any point in time as "cooperative" if $p \geq 0.8$. Likewise, we classify the state of the artificial society at any point in time as "compliant" if $c \geq 0.5$. These are entirely arbitrary values, but they are easily subject to sensitivity analysis to ensure that any qualitative conclusions we may find are not highly dependent on these modeling choices.

The two distinct hypotheses we examine in this paper are as follows:

H_A = The convict code tends to break down as prisoner heterogeneity increases.
H_B = The convict code tends to break down as the number of prisoners increases.

Our artificial society is a distillation of reality, and as such, we must rely on surrogates for the existence of a norm akin to the convict code. Therefore, the hypotheses we actually test with our model are as follows:

$H_{A'}$ = Measured levels of p (or c) tend to decrease as agent heterogeneity increases.
$H_{B'}$ = Measured levels of p (or c) tend to decrease as the number of agents increases.

To test these hypotheses, we first enable our artificial society to achieve a cooperative and compliant state with homogenous agents. Then, we shock the population with respect to the number of prisoners, the heterogeneity of prisoners, or both, and measure the extent to which the shocks have diminished the willingness or ability of agents to restore the cooperative outcome.

We consider two modes for introducing changes in the composition of the population: *swap* and *add*. The swap mode allows us to randomly identify agents of the initial/original population and replace them with new agents. The new agents can be identical to the prisoners they replace in terms of their *tagString*, they can be completely different, or they can be some variation in between. For the case described in this section, each time a swap occurs, 50 agents are swapped.

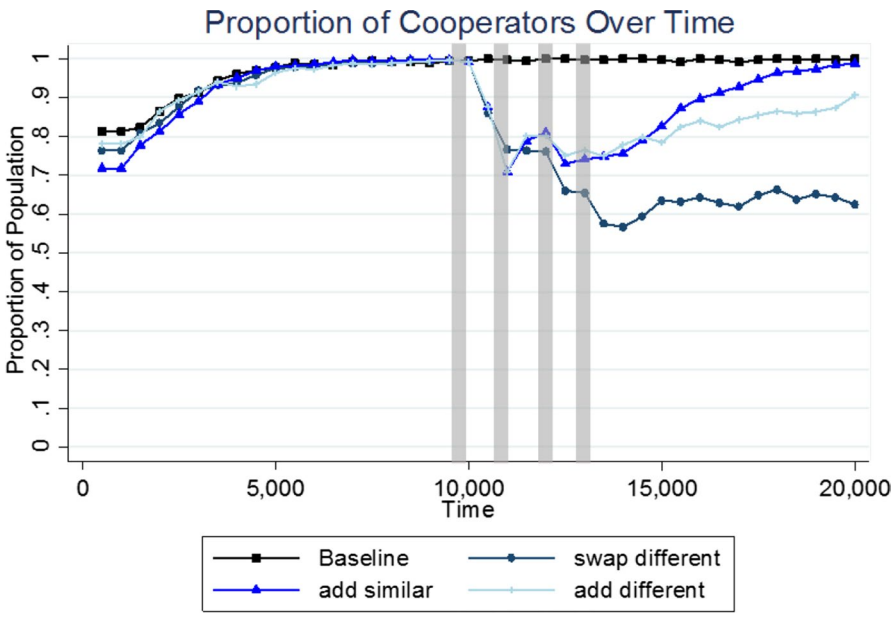

Fig. 4 Proportion of cooperators (p) over time for demonstration

The add mode allows us to introduce new agents into the population. As with the swap mode, the new agents can have identical *tagString*s or be completely different and each time an add occurs, 50 agents are added.

We examine four design points with one replication each. The first is our experimental control, in which we allow the model to run for all 20,000 time-steps with no treatments. For the second design point, we swap different agents. That is, at the specified times, 50 original agents (with tagString [1 1 1 …1]) are swapped out with new and different agents (tagString [0 0 0 … 0]). We compare this treatment with our control to test the heterogeneity hypothesis ($H_{A'}$). The third design point adds similar agents at the specified times. We compare this treatment with our control to test the prisoner increase hypothesis ($H_{B'}$). Finally, the fourth adds *different* agents at the specified times. This treatment tests a conceptual combination of the hypotheses. It is important to note that the strategy for any new agent introduced in the system is randomly assigned.[7]

Figure 4 depicts the change in p over time for each of the three treatments and the baseline. The vertical gray lines correspond to the treatment times. We first observe that the initial parameters are ones for which extremely high levels of cooperation are easy to obtain. With respect to the treatments, all treatments appear to at least temporarily diminish or undermine the ability or willingness for agents to cooperate.

[7] The strategies are shown in Table 2 on page 10. Since we assume agents are not hypocrites, this means that new agents are actually more likely to be cooperative, since there are relatively more cooperative strategies.

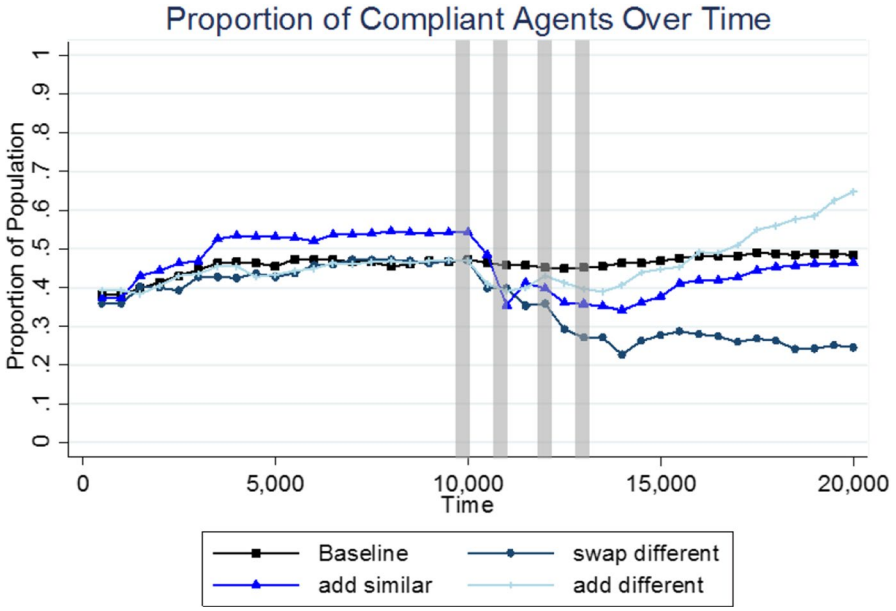

Fig. 5 Proportion of compliant (c) agents over time for demonstration

For example, at the time of the last treatment (time-step 13,000), no treatment group has greater than 75% cooperation proportion and cooperation proportion for the swap treatment is as low as 65%.

The effect of adding similar agents (medium blue line) appears to be the weakest and least persistent, as high levels of cooperation are restored before the end of the simulation for that design point. It also looks as though adding different agents (light blue line) has a relatively transient effect, though slightly larger in magnitude than adding similar agents. In contrast, swapping different agents has the largest effect and it appears to operate for a long period of time. Notice how the levels of p never regain its previous levels (near 100%) or even reach the threshold for a cooperative outcome ($>80\%$).We next turn our attention to compliant agents. Figure 5 depicts the change in c over time for each of the three treatments.

We observe that the proportion of compliant agents in the baseline group is just shy of 50% and the variance in this measure of effectiveness is greater before the treatment begins. As with p, the treatments all appear to negatively affect c at least temporarily. Again, the effect of adding agents appears to have a slightly smaller effect and also lasts a shorter period of time. In fact, adding different agent appears to jostle the model to a point where it is able to achieve higher levels of compliant agents than even the baseline achieves.

Again, the swap-different agents mode appears to have a larger and longer lasting effect. The proportion of compliant agents drops from a pre-treatment level of ~45% down to a low of approximately 22% where it holds fairly steady in the mid-twenties.

It is important to note that the state of the model immediately following the last treatment and the extent to which the model seems to regain a steady-state both help

to shape the narrative we obtain from the analysis. For example, it is not overly surprising that swapping different agents into the population immediately lowers cooperation and compliance. However, it is helpful to examine whether the agents can overcome that shock and regain a cooperative outcome. In this scenario, the other treatments appear nearly as disruptive at first, but not as persistent. We flesh out this narrative more fully in Sect. 5 and explore its implications in Sect. 6.

We have identified an admittedly narrow set of sufficient conditions under which $H_{A'}$ and $H_{B'}$ are true. However, we do not know how sensitive these relationships are to arbitrary parameter levels, nor do we yet know how robust or generalizable they are. With a more systematic approach we can leverage computational power along with statistical analysis to gain greater insight into the phenomenon of cooperation among individuals in our artificial prison society.

5 Analysis of experiment reveals systematic relationships

The purpose of this section is to rigorously measure the effect of modifying the demographics of our artificial prison population on agents' willingness to cooperate with others. We implement a full-factorial experimental design to create a response surface for each of the measures of effectiveness we consider and then fit various ordinary least squares (OLS) regression models to those surfaces in an effort to assess the nature of the treatment effects. The experimental design enables us to attribute a causal effect to the treatments.

We first identify eight design points, which we name Starting Design Points that have proven to achieve cooperative outcomes after 20 generations. These design points and the relevant factors associated with them are shown in Table 4. The choice of the starting points is perhaps the most challenging part of the experimental design. Cooperative outcomes, we have found, are relatively rare but seemingly robust when they do occur. Therefore, it is difficult to vary the parameters in Table 4 more widely, because in order for our subsequent analysis to work, we must be reasonably assured of achieving cooperation to begin with.

In our limited demonstration in the previous section, we employ four treatments with two generations between treatments and involve 50 agents in each treatment. In order to examine the sensitivity of our conclusions to these and other arbitrary decisions, we vary the factors in Table 5 in a full-factorial design. As the table shows, we consider four treatment combinations, and a wide range of values for arbitrary factors like number of treatments and treatment interval.[8] The design provides 1,672 unique design points, which we replicate 20 times each for a total of 33,400 runs. We find that 20 replications give our subsequent regression models sufficient statistical power to draw relevant conclusions.

The factor *Treatment Proportion* is the only factor we have not yet defined. This is the total number of agents involved in a complete treatment cycle for that design point as a proportion of initial number of agents (*Number of Inmates* for

[8] The treatments are the same as those described on page 17, above.

Table 4 Starting Design Points

Starting_DP	Number Inmates	payoff_T	payoff_R	payoff_P	payoff_S	payoff_D	payoff_E	witnessRange	metaWitnessRange
0	50	5	4	1	0	−6	−2	5	2
1	50	9	6	1	0	−12	−2	5	2
2	250	3	2	1	0	−12	−2	5	5
3	250	5	2	1	0	−6	−2	5	2
4	250	11	6	1	0	−6	−2	5	5
5	250	13	8	3	0	−12	−2	5	5
6	450	3	2	1	0	−12	−2	2	5
7	450	9	8	3	0	−12	−2	5	5

Table 5 Experimental design

Factor	Description	Type	Levels
Starting Design Point	Cooperative starting point	Categorical	[1, …, 8]
Mode	Type of treatment	Categorical	[add, swap]
Different	Are new agents different?	Binary	[yes, no]
Number of Treatments	Number of treatments	Integer	[1, 2, 3, 4, 5]
Treatment Interval	Number of generations between treatments	Integer	[1, 2, 3]
Treatment Proportion	Proportion of original population involved in complete treatment	Continuous	[25, 50, 75, 100]

that particular design point). If a design point starts with 250 agents and calls for a *Treatment Proportion* of 50, then the total number of agents involved in the treatment is $0.5 \cdot 250 = 125$. If the design point calls for one treatment, the treatment consists of 125 agents. Alternatively, if the design point calls for, say, 5 treatments, then each treatment consists of 25 agents.

In the following analysis, let d refer to the design point (1 to 1672) and r be the replication (1 to 20). Then, the proportion of cooperative agent after g generations is given by: $p_{dr}(g)$. Similarly, the proportion of compliant agents after g generations is given by: $c_{dr}(g)$.

For all design points treatments commence at 10,000 time-steps, or $g = 20$ generations. That is, upon completion of the twentieth generation, the set of new agents are either added or swapped. The simulation duration for each design point may differ because the number of treatments and the interval between treatments may differ. In all cases, we collect data for 20 generations after the final treatment. We define f_d as that number of generations at which the final treatment is given for design point d. So, for the example in Fig. 5 (page 20), $f = 26$, because the final treatment is applied at time $t = 13{,}000$, or after 26 generations. Finally, we define h as the count of the number of generations after the final treatment for that design point is implemented. We choose to start treatment after 20 generations and end the model 20 generations after the final treatment is given because initial analysis reveals that under most circumstances *if* a cooperative outcome were to occur, 20 generations was sufficient to achieve it.

Among the many benefits of simulation modeling is the ability to compare treatment outcomes to counterfactual outcomes over time. We have the luxury in this case of being able to run the model from a particular starting point and not implement any treatments. Thus, we can compare the proportion of compliant agents after treatments with the same time in an identical artificial society without treatment. In our notation, the bar indicates that the measure comes from the baseline output (cooperative starting point) and s may take on values 1 to 8.

We use the Greek letter β as a prefix which indicates the given response variable is the difference in proportion relative to the appropriate baseline. Thus, the difference in the proportion of cooperation x generations after the last treatment, relative to the baseline, is given by:

Table 6 Effect magnitudes of one generation after final treatment (with controls for Starting Design Point)

Factor	Response E[p'(1)]	Response E[c'(1)]
Number of Treatments	0.0265***	0.0142***
	(−0.00038)	(0.00032)
Treatment Interval	0.0452***	0.0212***
	(0.000512)	(0.00043)
Treatment Proportion	−0.044***	−0.0026***
	(1.66e−5)	(1.4e−5)
Add same agents	−0.143***	−0.0664***
	(0.007)	(0.0058)
Add different agents	−0.146***	−0.0335***
	(0.007)	(0.0058)
Swap same agents	−0.258***	−0.126***
	(0.007)	(0.0058)
Swap different agents	−0.330***	−0.152***
	(0.007)	(0.0058)
N	33,400	33,400
Adj *R*^2	0.785	0.639
Mean response	−0.321	−0.172

***$p<0.001$, **$p<0.01$, *$p<0.05$

$$\beta p_{dr}(x) = p_{dr}\left(f_d + x\right) - \overline{p}_{sr}\left(f_d + x\right)$$

where $\overline{p}_{sr}$ is the rth replication of the Starting Design Point that is associated with design point d.

We first examine the state of our artificial society immediately after treatment. We measure the proportion of cooperation and proportion of compliant agents one generation after the final treatment. For each of these response variables, we develop an ordinary least squares regression model with the factors listed in Table 5 (page 23) as the covariates. The regression models are listed in Table 6.[9]

The first regression has the difference in proportion of cooperative agents ($\beta p(1)$) as the response variable, and the next has the difference in proportion of compliant agents ($\beta c(1)$) as the response variable. The first set of parameter estimates relate to the number, frequency, and intensity of the treatments, while the next set is the estimates for the treatment modes.

The regression analysis confirms what the initial case (Sect. 4) suggests. Swapping out different agents or adding similar agents has a substantially negative effect on the proportion of cooperative and compliant agents. All else equal, swapping out different agents reduces the proportion of cooperative agents approximately 33.0 percentage points, while adding similar agents reduces that proportion by 14.3 percentage points, relative to the behavior of the untreated baseline. Similarly, swapping different agents reduces compliant agents by 15.2 percentage points if

[9] See Law and Kelton (2000) for more on regression analysis of simulation models.

Table 7 OLS regressions of $\beta p(5)$ to $\beta p(20)$ (with controls for Starting Design Point)

Factor	E[p'(5)]	E[p'(10)]	E[p'(15)]	E[p'(20)]
Number of Treatments	0.0096***	0.0110***	0.0085***	0.0088***
	(0.00041)	(0.00058)	(0.00061)	(0.00070)
Treatment Interval	0.0203***	0.0156***	0.0136***	0.0146***
	(0.00054)	(0.00077)	(0.00082)	(0.00093)
Treatment Proportion	−0.0029***	−0.0029***	−0.0022***	−0.0023***
	(1.8e−5)	(2.5e−5)	(2.6e−5)	(0.00003)
Add same agents	−0.0268**	0.0192	0.0199	0.0393**
	(0.0074)	(0.01)	(0.011)	(0.013)
Add different agents	−0.0348***	0.0177	0.0206	0.0422**
	(0.0074)	(0.01)	(0.011)	(0.013)
Swap same agents	−0.118***	−0.0689***	−0.0537***	−0.0410**
	(0.0074)	(0.01)	(0.011)	(0.013)
Swap different agents	−0.173***	−0.139***	−0.103***	−0.0972***
	(0.0074)	(0.01)	(0.011)	(0.013)
N	33,400	33,400	33,400	33,400
Adj *R*^2	0.588	0.444	0.312	0.292
Mean response	−0.201	−0.158	−0.114	−0.100

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

measured from the untreated baseline, and adding similar agents reduces the proportion of compliant agents by 6.6 percentage points.

Another important relationship is found in the *Treatment Proportion* parameter. While the coefficient for this parameter appears small, the factor level ranges from 25 to 100. So, when the number of agents involved in the treatment is as many as the initial population of agents (i.e., when *Treatment Proportion* = 100), then the effect from that parameter tends to reduce the proportion of cooperative agents by approximately 40 percentage points and to reduce the proportion of compliant agents by approximately 30 percentage points. In contrast, we see that the more arbitrary factors, *Number of Treatments* and *Treatment Interval* have small magnitudes relative to the other factors. For example, only at its maximum level of 5 is the effect from *Number of Treatments* as large as 13 percentage points.

In the above section, we find strong evidence that our treatments cause a substantial reduction in the proportion of cooperative and the proportion of compliant agents in our artificial society. In this section, we examine the extent to which we may expect these effects to persist. We do so by creating regression models of the difference in the proportion of cooperative agents and proportion of compliant agents through time after final treatment. The OLS regressions for $\beta p(5)$ to $\beta p(20)$ are found in Table 7.

The magnitude of the negative effect of swapping different agents is persistent over time and relatively large. The drop is approximately 17.3 percentage points after five generations and remains as large as 9.7 percentage points after twenty. After five generations, the magnitude of the effect of adding similar agents is only

Table 8 OLS regressions of $\beta c(5)$ to $\beta c(20)$ (with controls for Starting Design Point)

Factor	E[c'(5)]	E[c'(10)]	E[c'(15)]	E[c'(20)]
Number of Treatments	0.0108***	0.0078***	0.0070***	0.0059***
	(0.00036)	(0.00046)	(0.00056)	(0.00063)
Treatment Interval	0.0185***	0.0109***	0.0104***	0.0107***
	(0.00048)	(0.00061)	(0.00074)	(0.00084)
Treatment Proportion	−0.0020***	−0.0020***	−0.0017***	−0.0017***
	(1.6e−5)	(2.0e−5)	(2.4e−5)	(2.7e−5)
Add same agents	−0.0449***	0.0171*	0.0226*	0.0355**
	(0.0065)	(0.0083)	(0.01)	(0.011)
Add different agents	0.0108	0.106***	0.147***	0.188***
	(0.0065)	(0.0083)	(0.01)	(0.011)
Swap same agents	−0.0922***	−0.0277**	−0.0151	−0.0041
	(0.0065)	(0.0083)	(0.01)	(0.011)
Swap different agents	−0.0937	−0.0239*	0.0172*	0.0363*
	(0.0065)	(0.0083)	(0.01)	(0.011)
N	33,400	33,400	33,400	33,400
Adj *R*^2	0.49	0.465	0.425	0.418
Mean response	−0.111	−0.061	−0.021	0.000

***$p<0.001$, **$p<0.01$, *$p<0.05$

approximately −2.7% and remains small as the number of generations increases, and even turns slightly positive. As before, the effect of *Treatment Proportion* is negative, relatively large, and persistent over each model. The effect of *Number of Treatments* and *Treatment Interval* also appears minor.

Next, we examine the same scheme of regression models, but in this case we focus on the proportion of compliant agents. Table 8 contains those response variables measured from the untreated baseline.

Swapping different agents has a negative effect on the proportion of compliant agents in the population after five generations (−9.4 percentage points), but by ten generations the effect is not practically significant and even turns positive by fifteen. Adding similar agents has a modest negative effect on compliant agents (−4.5 percentage points) at five generations, but by ten generations, and out to twenty generations, the effect is positive but small. Once again the effect of *Treatment Proportion* is substantial.

In this section, we begin our analysis by initializing the model with starting points confirmed to achieve a cooperative result. The mechanisms in place, namely punishment of norm violators; punishment of those who fail to punish norm violators; and observable tagStrings, all help to stack the deck in favor of the cooperative outcome. However, we confirm that swapping different agents tends to have a persistent and substantial effect on the proportion of cooperative agents, even after as many as twenty generations. The effect of swapping different agents appears to only negatively affect compliant agents for five or ten generations. Adding agents tends to negatively and modestly affect the number of cooperative and compliant agents, but

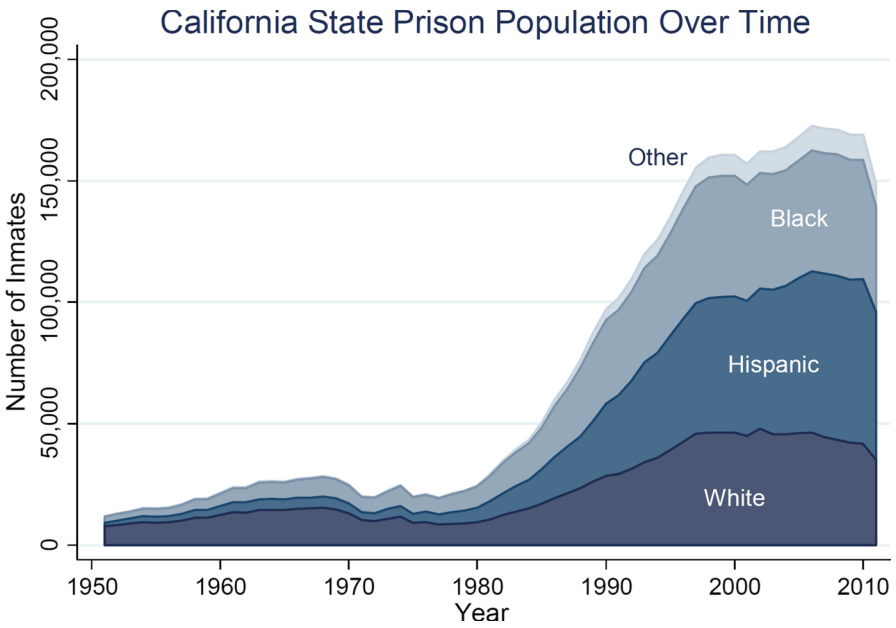

Fig. 6 Time-plot of California prison population

this effect tends to persist to only five generations at most. The most robust finding of all is that the proportion of agents involved in the treatments relative to the number of original agents is highly negative and of high practical significance.

6 Calibration: social order in the California prison system

From 1951 to 2006, the prison population in the state of California increased by a factor of 14 and ethnic fractionalization increased by 30% (Skarbek 2014). The experimental design we describe above demonstrates substantial effects of relatively modest changes in the population. In this section, we briefly outline a demonstration of the model with more dramatic, and more realistic, population changes.

For this analysis, we initiate the model with 100 inmate agents and when it comes time to apply treatments, we do so in a manner that maintains the relative growth (or reduction) of agents in each ethnic category. We initiate our population with 100 inmate agents for two reasons. The first is that a 14-fold increase is still easily managed by the model from that starting point, while a model that includes the full population of approximately 170,000 inmates is not. The second reason is that modeling the entire California prison population is not necessary because that population was spread across dozens of prison; thus, not all inmates had the opportunity to interact with all other inmates.

Two periods of change that this population experienced are important to point out at this time. For the first twenty years, we see a slow but steady increase in the

Table 9 Inter-type similarity levels

Similarity	White	Black	Hispanic	Other
White	1.000	0.000	0.467	0.533
Black	0.000	1.000	0.533	0.467
Hispanic	0.467	0.533	1.000	0.000
Other	0.533	0.467	0.000	1.000

overall prison population, such that the totals we observe in the early 1970s are nearly triple that of 1951. Second, beginning in the late 1970s, we see explosive growth for the next two decades. Historically, the decline of the convict code is associated with this period (Skarbek 2014). See Fig. 6 for a graphical depiction of how the prison population changed during this time.

We initiate the model with 63 white agents, 15 Hispanic agents, 20 black agents, and 2 "other" agents.[10] We differentiate ethnic groups with different, immutable tag-Strings. Table 9 displays the relative similarities between each group.

Recall that *similarity_threshold* is a parameter that determines the extent to which agents treat each other as similar or different. If an agent pair's calculated similarity (see Table 9) exceeds the threshold, the agents consider each other sufficiently "similar" or fellow insiders. Thus, if the similarity threshold is set at a level greater than 0.534, none of the groups in our society see members of other groups as similar. In the analysis that follows, *similarity_threshold* is set to 0.8, which essentially renders all groups to see others as outsiders. This is generally consistent with observations of present-day prison societies, though the qualitative conclusions are robust for values greater than zero.

In order to replicate the demographic changes that the California prison population experienced during this period, we apply a sequence of treatments to our artificial society. After a transient period of 20 generations in order to allow for a spontaneous cooperative outcome to occur (as observed in reality by Irwin 1980), we swap and add agents so as to achieve relative growth rates for each ethnic group that are identical to the empirical annual rates. We simply add (or subtract) agents of each type at the rate described in the data. We swap out approximately 5% of the population in an ethnic-neutral fashion during each round of treatment in order to help account for the fact that inmates are occasionally released from prison.[11] In this manner, we achieve both the explosive growth in the prison population while matching the ethnic fractionalization over time.

The primary modeling challenge is to decide upon the rate at which rounds of treatments should occur. Ultimately, this gets to the question of the relationship between a time-step in the model and units of time in reality. In the results we present below, we decide upon 60 rounds of treatment because of the 60 years between 1951 and 2011. We choose a treatment interval of 1 generation (500

[10] The California Department and Corrections classifies non-white, non-black, and non-Hispanic prisoners into the category "other."

[11] The qualitative results from this section are similar if this parameter is set to zero.

Table 10 Parameter values for calibrated runs

Parameter	Value
Number of Inmates	100
Size of tagString	15
Generation Duration	500
Payoff T	9
Payoff R	4
Payoff P	1
Payoff S	0
Payoff D	−12
Payoff E	−2
Witness range	5
Meta-witness range	5
Similarity threshold	80

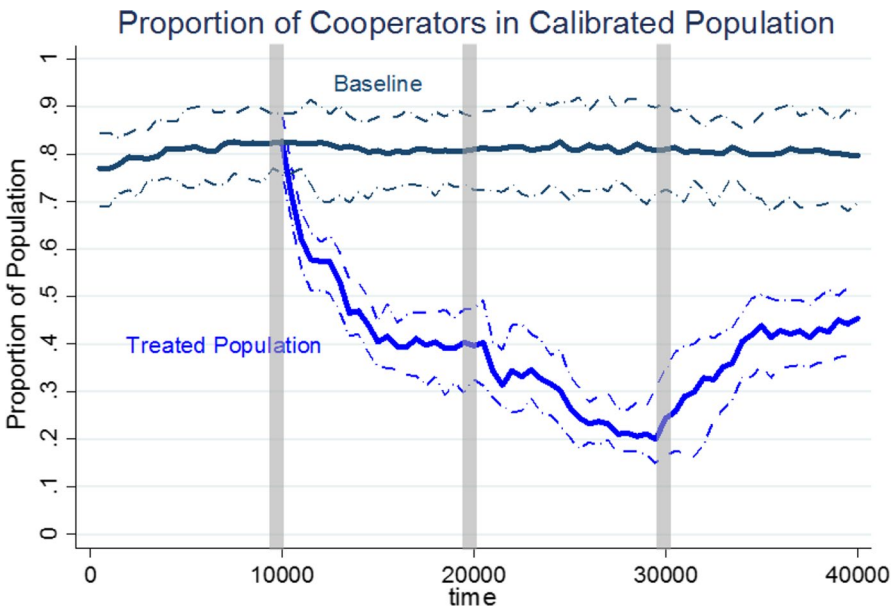

Fig. 7 Cooperators over time in calibrated population. Each datapoint on either main line is the mean of 30 replications at that time. The dotted lines are 90% empirical confidence intervals. The vertical gray bars depict the first, the twentieth, and the fortieth treatments

time-steps), which means that at the end of each generation (after a warm-up of 20 generations or 10,000 time-steps) we apply another treatment where we add and swap agents. This matching of rounds to years is admittedly arbitrary; however, the qualitative outcomes are robust for treatment intervals as large as 5.

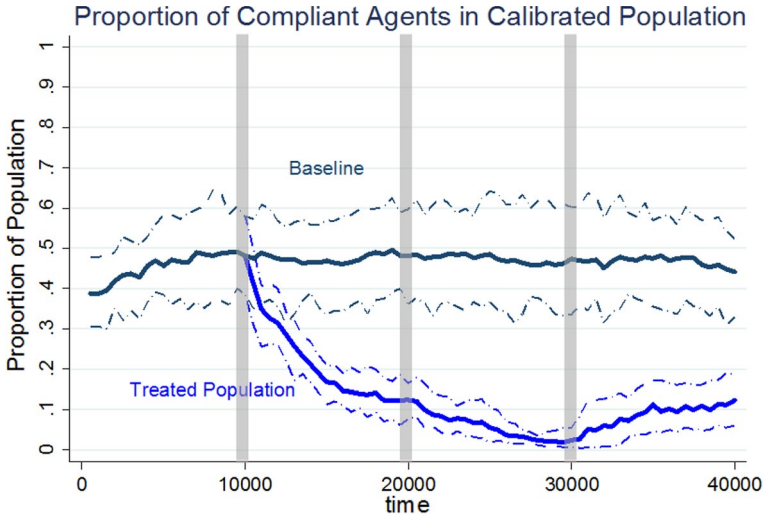

Fig. 8 Compliant agents over time in calibrated population. Each datapoint on a main line is the mean of 30 replications at that time. The dotted lines are 90% empirical confidence intervals. The vertical gray bars depict the first, the twentieth, and the fortieth treatments

The additional relevant parameters for the results we display are shown in Table 10. The primary reason we select these parameters is because 100 agents operating under the conditions shown below will tend to spontaneously achieve a cooperative outcome (i.e., proportion of cooperative agents in population > 0.8 or proportion of compliant agents > 0.5).

We implement the given parameters with the treatment scheme described above and replicate the model 30 times. For control purposes, we also run a baseline for each replication with no treatments but with common random numbers so that the only difference between the treatment replication and its respective baseline replication is the treatment scheme.[12]

Figure 7 illustrates that to radically change the number and demographic distribution of agents in our artificial society dramatically reduces the tendency of agents to select cooperative strategies. The first vertical gray line is at 10,000 and highlights the first treatment. We see the proportion of cooperative agents drops precipitously and nearly monotonically for approximately 40 treatments. Notice that the propensity to cooperate is essentially cut in half by treatment 20, which corresponds to 1971. By treatment 41 (1991), the proportion of cooperators has reached its minimum of approximately 20% and never fully recovers.

We see a similar narrative in Fig. 8. The proportion of agents who select compliant strategies drops after the first treatment and continues to drop steadily for the next forty or so rounds. At treatment 21 (approximately 1971), we observe

[12] See Law and Kelton (2000: pp 582–584) for more information on the use of common random numbers as a variance reduction technique.

compliant rates of approximately 15% in the treatment population, while the baseline population enjoys levels of nearly 50%. The proportion of compliant agents in the population bottoms out by treatment 41 and never comes close to recovering by the end of the experiment.

This simulation shows how our model can replicate the dramatic demographic changes in an actual prison population and how it influences the observance (Irwin 1980) and then breakdown (Hunt et al. 1993; Skarbek 2014) of the convict code. After only 20 rounds of treatments, which corresponds to 20 years of moderate growth in the actual prison population under consideration, we observe a precipitous decline in the proportion of agents that select compliant or even cooperative strategies. This result is robust with regard to arbitrary modeling decisions such as the treatment interval and degree of similarity between modeled ethnic groups. This section provides evidence in favor of the view that the convict code could not endure the stress of these demographic changes in the population therefore disintegrated.

7 Discussion

The emergence of social order in the absence of strong, effective third-party enforcement is often fraught with failure. Past work finds that norms can play a key role in promoting social cooperation, but only in some settings (Ostrom 1990; Ellickson 1991). The ability to choose with whom to interact has been a key reason for self-enforcing exchange to lead to desirable outcomes (Stringham 2015). In this paper, we remove the ability for this mechanism to operate by modeling the interaction of prisoners who cannot opt out of interactions with other agents.

We construct an agent-based model to test two hypotheses that relate to the evolution of norms within a prison society. The first is increasing the heterogeneity of the agents in our artificial society tends to decrease levels of cooperation and compliance. The second is increasing the number of agents in our society tends to decrease levels of cooperation and compliance. We find strong evidence in favor of both of these hypotheses that are robust to a wide variety of input parameters. Swapping out different agents tends to have an immediate and negative effect on rates of cooperation and compliance that are large and relatively persistent. Adding similar agents also tends to have an immediate, negative effect, though not as large or quite as persistent. In addition, we use empirical data from the state of California to calibrate our model with respect to rate of population increase and increase in heterogeneity. We demonstrate how the convict code may have disintegrated under the stress of these dramatic demographic changes. Our findings are robust to a number of different modeling decisions.

The paper also provides an opportunity to understand better the dynamics of norm emergence and their implications for prison management. The USA incarcerates a larger number and rate of its population than other country, driven partly by longer sentences, more aggressive prosecutors, and the increasing level of public punitiveness (Enns 2014). This large prison population stands in stark contrast to western European countries. Moreover, these countries also have significantly different informal institutions within their prisons and, in particular, lack racially

segregated prison gangs like those that exist in California (Skarbek 2016). If prison gangs emerge when norm-based governance fails, then mass incarceration carries even greater costs than previously believed. These costs include enhanced criminal activity within prisons and an increase in the recidivism rates of gang members (Dooley et al. 2014). Improvements could be made if prison populations were better managed in smaller and safer facilities. Our model formalizes an intuition about the role of changes in prisoner demographics that provides a reason for greater skepticism about current incarceration levels. In doing so, it links together the relationship about how changes in formal institutions—officials' decisions about prison size and management practice—cause changes in informal institutions—the importance of norm-based governance among prisoners.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11403-021-00316-7.

References

Akerlof GA, Kranton RE (2000) Economics and identity. Quart J Econ 115(3):715–753
Akerlof GA, Kranton RE (2005) Identity and the economics of organizations. J Econ Perspect 19(1):9–32
Austin J, Smith E, Srinivasan S, Sanchez F (2012) Social dynamics of gang involvement: a mathematical approach. Working paper
Avio K (1998) The economics of prisons. Eur J Law Econ 6(2):143–175
Axelrod R (1986) An evolutionary approach to norms. Am Polit Sci Rev 80(4):1095–1111
Bendor J, Mookherjee D (1990) Norms, third-party sanctions, and cooperation. J Law Econ Organ 6(1):33–63
Benjamin DJ, Choi JJ, Strickland AJ (2010) Social identity and preferences. Am Econ Rev 100(4):1913–1928
Bowker LH (1980) Prison victimization. Elsevier Science Ltd, Amsterdam
Carpenter J, Matthews PH (2009) What norms trigger punishment? Exp Econ 12(3):272–288
Carpenter JP, Matthews PH (2010) Norm enforcement: the role of third parties. J Inst Theor Econ 166(2):239–258
Chen R, Chen Y (2011) The potential of social identity for equilibrium selection. Am Econ Rev 101(6):2562–2589
Clemmer D (1940) The prison community. Christopher Publishing, Boston
Cubitt RP, Drouvelis M, Gächter S (2011) Framing and free riding: emotional responses and punishment in social dilemma games. Exp Econ 14(2):254–272
De Cremer D, Van Vugt M (2002) Intergroup and intragroup aspects of leadership in social dilemmas: a relational model of cooperation. J Exp Soc Psychol 38(2):126–136
De Marchi S, Page SE (2014) Agent-based models. Annu Rev Polit Sci 17:1–20
DiIulio JJ (1996) Help wanted: economists, crime, and public policy. J Econ Perspect 10(1):3–24
Dixit A (2004) Lawlessness and economics: alternatives modes of governance. Princeton University Press, Princeton
Dooley BD, Seals A, Skarbek D (2014) The effect of prison gang membership on recidivism. J Crim Justice 42(3):267–275
Ellickson R (1991) Order without law: how neighbors settle disputes. Harvard University Press, Cambridge
Enns PK (2014) The public's increasing punitiveness and its influence on mass incarceration in the United States. Am J Polit Sci 58(4):857–872
Epstein JM (2002) Modeling civil violence: an agent-based computational approach. In: Proceedings of the National Academy of Sciences in the United States of America, vol 99, no 3
Fehr E, Gachter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90(4):980–994

Fehr E, Fischbacher U (2004) Third-party punishment and social norms. Evol Hum Behav 25(2):63-87
Fischer I (2009) Friend or foe: subjective expected relative similarity as a determinant of cooperation. J Exp Psychol Gen 138(3):341
Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54(3):533–554
Goette L, Huffman D, Meier S (2006) The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. Am Econ Rev 96(2):212–216
Goh CK, Quek K, Tan KC, Abbass HA (2006) Modeling civil violence: an evolutionary multi-agent, game theoretic approach. IEEE Congress on Evolutionary Computation, July 16–21, 2006
Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am Econ Rev 83:525–548
Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, Cambridge
Hales D (2001) Cooperation without memory or space. Lect Notes Comput Sci 1979:157–166
Haney C, Banks C, Zimbardo P (1972) Interpersonal dynamics in a simulated prison (No. ONR-TR-Z-09). Stanford University, Department of Psychology
Holland J (1993) The effects of labels (tags) on social interactions. Working paper, Sante Fe Institute, 93-10-064
Horne C (2001) The enforcement of norms: group cohesion and meta-norms. Soc Psychol Quart 64(3):253–266
Hunt G, Riegel S, Morales T, Waldorf D (1993) Changes in prison culture: prison gangs and the case of the 'Pepsi generation.' Soc Probl 40(3):398–409
Irwin J (1980) Prisons in Turmoil. Little, Brown, & Co, Boston
Irwin J, Cressey D (1962) Thieves, convicts, and the inmate culture. Soc Probl 10(2):142–155
Janssen M (2008) Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. J Econ Behav Organ 65:458–471
Kendal J, Feldman MW, Aoki K (2006) Cultural coevolution of norm adoption and enforcement when punishers are rewarded or non-punishers are punished. Theor Popul Biol 70(1):10–25
Kusakawa T, Ogawa K, Shichijo T (2012) An experimental investigation of a third-person enforcement in a prisoner's dilemma game. Econ Lett 117(3):704–707
Law AM, Kelton WD (2000) Simulation modeling and analysis, 3rd edn. McGraw Hill, Boston
Leeson PT (2007a) An-arrgh-chy: the law and economics of pirate organization. J Polit Econ 115(6):1049–1094
Leeson PT (2007b) Efficient anarchy. Public Choice 130(1–2):41–53
Leeson PT (2009) The laws of lawlessness. J Legal Stud 38(2):471–503
Leeson PT (2010) Pirational choice: the economics of infamous pirate practices. J Econ Behav Organ 76(3):497–510
Leeson PT (2014) Anarchy unbound: why self-governance works better than you think. Cambridge University Press, Cambridge
Luther WJ (2015) The monetary mechanism of stateless Somalia. Public Choice 165(1–2):45–58
Mahmoud S, Griffiths N, Keppens J, Luck M (2012) Norm emergence through dynamic policy adaptation in scale free networks. In: International workshop on coordination, organizations, institutions, and norms in agent systems. Springer, Berlin, Heidelberg
Melleson N, Heppenstall A, See L (2012) Crime reduction through simulation: an agent-based model of burglary. Comput Environ Urban Syst 34(3):236–250
Mildenberger CD (2015) Virtual world order: the economics and organizations of virtual pirates. Public Choice 164(3–4):401–421
Mitchell MM, Fahmy C, Pyrooz DC, Decker SH (2016) Criminal crews, codes, and contexts: differences and similarities across the code of the street, convict code, street gangs, and prison gangs. Deviant Behav 38:1–26
Murtazashvili I, Murtazashvili J (2015) Anarchy, self-governance, and legal titling. Public Choice 162(3–4):287–305
North MJ, Collier NT, Ozik J, Tatara E, Altaweel M, Macal CM, Bragen M, Sydelko P (2013) Complex adaptive systems modeling with repast symphony. In: Complex adaptive systems modeling. Springer, Heidelberg, FRG
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge

Orbell JM, Schwartz-Shea P, Simmons RT (1984) Do cooperators exit more readily than defectors? Am Polit Sci Rev 78(01):147–162
Powell B, Stringham EP (2009) Public choice and the economic analysis of anarchy: a survey. Public Choice 140(3–4):503–538
Powell B, Wilson BJ (2008) An experimental investigation of Hobbesian jungles. J Econ Behav Organ 66(3):669–686
Pratt TC, Cullen FT (2000) The empirical status of Gottfredson and Hirschi's general theory of crime: a meta-analysis. Criminology 38(3):931–964
Prietula MJ, Conway D (2009) The evolution of metanorms: Quis custodiet ipsos custodes? Comput Math Organ Theory 15(3):147–168
Radford RA (1945) The economic organisation of a POW camp. Economica 12(48):189–201
Riolo RL (1997) The effects of tag-mediated selection of partners in populations playing the iterated prisoner's dilemma. In: Proceedings of the international conference of genetic algorithms
Riolo RL, Cohen MD, Axelrod R (2001) Evolution of cooperation without reciprocity. Nature 414:441–443
Rodrik D (2015) Economics rules: why economics works, when it fails, and how to tell the difference. Oxford University Press, Oxford
Rudoff A (1964) Prison inmates: an involuntary association. Dissertation. University of California, Berkeley
Sampson RJ, Raudenbush SW, Earls F (1997) Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277(5328):918–924
Schuessler R (1989) Exit threats and cooperation under anonymity. J Confl Resolut 33(4):728–749
Skarbek D (2012) Prison gangs, norms, and organizations. J Econ Behav Organ 82:96–109
Skarbek D (2014) The social order of the underworld: how prison gangs govern the American penal system. Oxford University Press, Oxford
Skarbek D (2016) Covenants without the sword? Comparing prison self-governance globally. Am Polit Sci Rev 110(4):845–862
Skarbek D (2020) The puzzle of prison order: why life behind bars varies around the world. Oxford University Press, Oxford
Stringham EP (2015) Private governance: creating order in economic and social life. Oxford University Press, Oxford
Sykes GM (1958) The society of captives: a study of a maximum security prison. Princeton University Press, Princeton
Sykes G, Messinger SL (1962) The inmate social code and its functions. In: Johnston N, Savitz L, Wolfgang ME (eds) The sociology of punishment and correction. Wiley, New York
Tajfel H (1974) Social identity and intergroup behaviour. Inf Int Soc Sci Counc 13(2):65–93
Tajfel H, Turner J (1979) An integrative theory of intergroup conflict. In: Austin WG, Worchel S (eds) The social psychology of intergroup relations. Brooks/Cole Pub. Co., Monterey
Tako A, Robinson S (2010) Model development in discrete-event simulation and system dynamics: an empirical study of expert modelers. Eur J Oper Res 207(1):784–794
Tullock G (1985) Adam Smith and the prisoner's dilemma. Quart J Econ 100:1073–1081
Vanberg VJ, Congleton RD (1992) Rationality, morality, and exit. Am Polit Sci Rev 86(2):418–431
Williams VL, Fish M (1974) Convicts, codes, and contra-band: the prison life of men and women. Ballinger Publishing Company, Cambridge
Zou Y, Fonoberov V, Fonoberova M, Mezic I, Kevrekidis I (2012) Model reduction for agent-based social simulation: coarse-graining a civil violence model. Phys Rev 85:1–13