Centralized sanctioning institutions have been shown to emerge naturally through social learning, displace all other forms of punishment and lead to stable cooperation. However, this result provokes a number of questions. If centralized sanctioning is so successful, then why do many highly authoritarian states suffer from low levels of cooperation? Why do states with high levels of public good provision tend to rely more on citizen-driven peer punishment?

Here, we consider how corruption influences the evolution of cooperation and punishment. Our model shows that the effectiveness of centralized punishment in promoting cooperation breaks down when some actors in the model are allowed to bribe centralized authorities. Counterintuitively, a weaker centralized authority is actually more effective because it allows peer punishment to restore cooperation in the presence of corruption. Our results provide an evolutionary rationale for why public goods provision rarely flourishes in polities that rely only on strong centralized institutions. Instead, cooperation requires both decentralized and centralized enforcement. These results help to explain why citizen participation is a fundamental necessity for policing the commons.

A centuries-old debate exists on how to best govern society and promote cooperation: is cooperation best maintained by a central authority [1,2] or is it better handled by more decentralized forms of governance [3,4]? The debate is still unresolved, and identifying mechanisms that promote cooperation remains one of the most difficult challenges facing society and policymakers today [4].

Decentralized, individual sanctioning of non-cooperators (also known as free-riders or defectors) is one of the main tools used by societies to promote and maintain cooperation [5]. Individuals can sanction free-riders implicitly via behavioural reciprocity (as in the case of the highly successful tit-for-tat strategy [6]) or explicitly via costly punishment [7]. Both of these forms of peer punishment have been widely studied using evolutionary models and behavioural experiments [8–10,6,11].

Recently, however, Sigmund et al. [12] showed that centralized institutions can have an evolutionary advantage over peer punishment because, unlike peer-punishers, these institutions may eliminate ‘second-order’ free-riding. Second-order free-riders cooperate with other players but they do not pay the cost of punishing defectors and this can allow defectors to re-emerge [13–15]. To address this problem, Sigmund et al. present a model of ‘pool’ punishment, where agents commit resources to a centralized authority that sanctions free-riders [12,16]. Pool punishment avoids the second-order free-rider problem because the centralized authority punishes any individual who does not
contribute to the punishment pool (including cooperators and peer-punishers). This allows pool-punishers to quickly take over a population, displacing both free-riders and peer-punishers [12]. These advantages help to explain why human societies frequently delegate punishment to centralized institutions [12,17,16]. They also help to explain why centralized institutions acquire an increasing monopoly over legitimate punishment over time by stigmatizing and criminalizing [19, p. 371,372] various forms of peer punishment.

However, the dominance of pool punishment in the Sigmund et al. model [12] also creates three puzzles. First, the results imply that increasing the severity of centralized pool punishment always increases cooperation. Yet, many authoritarian states, which have the ability to severely punish citizens, suffer from low levels of participation and public goods provision [20]. Meanwhile, states with high levels of public goods, such as western democracies [20–22], typically limit the government’s ability to punish individuals and tolerate more forms of peer punishment.

Second, centralized pool punishment quickly takes over a population and completely displaces peer punishment [12] in the Sigmund et al. model [12], but many (if not most) societies exhibit a mix of centralized and decentralized punishment strategies. Even in societies with centralized punishment, citizens engage in costly acts of protest against agents who harm the public good. As recent events—from the Occupy protests to the Arab Spring—illustrate, this occurs even when the government punishes protesters [23–25]. What unmodelled factors might allow peer punishment to evolve alongside centralized enforcement institutions—even when these institutions are actively hostile towards various forms of peer punishment?

Third, the Sigmund et al. model [12] assumes that the centralized authority punishes all forms of peer punishment. This is because peer-punishers in their model, by definition, do not contribute to the centralized authority. However, many societies with centralized enforcement also recognize certain forms of peer punishment as legitimate. For instance, civil litigation, jury duty, anti-incumbent voting and other forms of peer punishment as legitimate. For instance, civil litigation, jury duty, anti-incumbent voting and other forms of peer punishment are also instances of altruistic peer punishment.

Here, we show that allowing for corruption in the model can help to explain how such hybrid strategies may evolve. Instead, cooperation rests on an authority that protects a fundamental aspect of civil society, citizen participation in policing the commons [29,30].

Our baseline model is a public good game (PGG) with both peer and pool punishments [12]. The PGG is a simple model for studying contributions to a project with non-excludable positive externalities, which may include everything from the provision of social insurance to the protection of the environment. Let $M$ denote the population size and let $N \leq M$ denote the number of individuals who are randomly chosen in a given round to play a PGG. In the game, each individual is faced with a choice: whether or not to contribute a fixed amount, $c > 0$, to the common pool. Once each individual chooses her action, each individual will obtain $rc(N/N_c)$, where $r$ is a factor greater than $1$, $N_c$ is the number of contributors to the common pool and $N$ is the total number of participants (whether they contributed or not). If all individuals contribute, $N_c = N$, then the social welfare is maximized and each individual obtains $r$. However, each actor gains an equal share of $rc(N/N_c)$, whether or not they contribute, making it a dominant strategy for each individual to free-ride by contributing $0$ (the pay-offs are written explicitly in the electronic supplementary material).

The population includes $X$ cooperators, who contribute $c$ and $Y$ defectors (free-riders), who do not. Consistent with previous work, we also assume that the game is not compulsory and some players may choose not to participate in the PGG [12,31–34]. These $Y$ defectors earn a fixed small pay-off, $\sigma$. In addition, $W$ peer-punishers cooperate by contributing $c$ to the PGG but also impose a fine, $\beta$, on each free-rider at a cost $\gamma$ [5]. In other words, each free-rider pays a total fine $\beta N_{p}$ where $N_p$ is the number of peer-punishers in the group, and every peer-punisher incurs an extra cost $\gamma N_p$, where $N_p$ is the number of free-riders in the group. Furthermore, peer-punishers inflict a penalty on cooperators proportional to the number of defectors (second-order punishment). We also have $V$ pool-punishers who, instead of directly punishing free-riders, contribute a fixed amount, $G$, to a punishment pool before participating in the game and then contribute $c$ to the PGG. Those who do not contribute to the pool (including free-riders, cooperators and peer-punishers) are then fined $BN_v$, where $N_v$ is the number of pool-punishers in the group. We also introduce $C$ corruptors to the model. A corruptor pays the central authority a fixed fee $KG$ to avoid being punished for not contributing to the PGG (this only makes sense if the fee is less than the total contributions paid by pool-punishers, $KG < G + c$).

We study the equilibria of fully mixed populations of fixed size $M$ and variable composition by computing the pay-offs obtained by players using these strategies, assuming that agents play in randomly sampled groups of size $N$. The difference in pay-offs, together with the parameter $s \geq 0$, determines the rate at which individuals with lower pay-offs are replaced by types with higher pay-offs. As in other evolutionary models, this process can be interpreted either as evolution or social learning. We also allow for random switching of strategies with a mutation rate $\mu \geq 0$. We derive equilibria as the long-run distribution of different strategies both analytically (in the limit of strong imitation) and using numerical simulation. For details, see the electronic supplementary material.

Figure 1 shows sample runs of a numerical simulation of the model. Without corruption, pool punishment eventually rarely flourish in polities with strong centralized punishment alone. Instead, cooperation rests on an authority that protects a fundamental aspect of civil society, citizen participation in policing the commons [29,30].

Our baseline model is a public good game (PGG) with both peer and pool punishments [12]. The PGG is a simple model for studying contributions to a project with non-excludable positive externalities, which may include everything from the provision of social insurance to the protection of the environment. Let $M$ denote the population size and let $N \leq M$ denote the number of individuals who are randomly chosen in a given round to play a PGG. In the game, each individual is faced with a choice: whether or not to contribute a fixed amount, $c > 0$, to the common pool. Once each individual chooses her action, each individual will obtain $rc(N/N_c)$, where $r$ is a factor greater than $1$, $N_c$ is the number of contributors to the common pool and $N$ is the total number of participants (whether they contributed or not). If all individuals contribute, $N_c = N$, then the social welfare is maximized and each individual obtains $r$. However, each actor gains an equal share of $rc(N/N_c)$, whether or not they contribute, making it a dominant strategy for each individual to free-ride by contributing $0$ (the pay-offs are written explicitly in the electronic supplementary material).

The population includes $X$ cooperators, who contribute $c$ and $Y$ defectors (free-riders), who do not. Consistent with previous work, we also assume that the game is not compulsory and some players may choose not to participate in the PGG [12,31–34]. These $Y$ defectors earn a fixed small pay-off, $\sigma$. In addition, $W$ peer-punishers cooperate by contributing $c$ to the PGG but also impose a fine, $\beta$, on each free-rider at a cost $\gamma$ [5]. In other words, each free-rider pays a total fine $\beta N_{p}$ where $N_p$ is the number of peer-punishers in the group, and every peer-punisher incurs an extra cost $\gamma N_p$, where $N_p$ is the number of free-riders in the group. Furthermore, peer-punishers inflict a penalty on cooperators proportional to the number of defectors (second-order punishment). We also have $V$ pool-punishers who, instead of directly punishing free-riders, contribute a fixed amount, $G$, to a punishment pool before participating in the game and then contribute $c$ to the PGG. Those who do not contribute to the pool (including free-riders, cooperators and peer-punishers) are then fined $BN_v$, where $N_v$ is the number of pool-punishers in the group. We also introduce $C$ corruptors to the model. A corruptor pays the central authority a fixed fee $KG$ to avoid being punished for not contributing to the PGG (this only makes sense if the fee is less than the total contributions paid by pool-punishers, $KG < G + c$).

We study the equilibria of fully mixed populations of fixed size $M$ and variable composition by computing the pay-offs obtained by players using these strategies, assuming that agents play in randomly sampled groups of size $N$. The difference in pay-offs, together with the parameter $s \geq 0$, determines the rate at which individuals with lower pay-offs are replaced by types with higher pay-offs. As in other evolutionary models, this process can be interpreted either as evolution or social learning. We also allow for random switching of strategies with a mutation rate $\mu \geq 0$. We derive equilibria as the long-run distribution of different strategies both analytically (in the limit of strong imitation) and using numerical simulation. For details, see the electronic supplementary material.

Figure 1 shows sample runs of a numerical simulation of the model. Without corruption, pool punishment eventually
takes over the population [12], and does so even earlier when $B$, the severity of second-order pool punishment, is higher.

The situation changes dramatically when we introduce corruptors. Pool punishment is no longer a dominant strategy, as shown in a sample simulation run (figure 1c). Interestingly, figure 1d shows that weakening pool punishment (lowering the fine $B$) allows peer-punishers to re-emerge as a relatively stable strategy that restores cooperation in the presence of corruption.

We investigate this further in figure 2a, which shows the proportion of different strategies as a function of second-order punishment severity. For low values of $B$, peer-punishers dominate and prevent the corruptor strategies from gaining ground. As $B$ increases, peer-punishers disappear and pool-punishers...
become more prevalent. However, with even higher values of $B$, the prevalence of corruptors also increases. This causes the total number of cooperative individuals to decline. We confirm these results by analytical computation of the long-run frequencies of strategies in the $(X, Y, Z, V, W, C)$ subpopulation (for methods, see the electronic supplementary material). With low $B$, the frequencies, respectively, are $1/(M + 7)$, confirming the clear dominance of peer-punishers (with a population of $M = 100$, this is approx. $[0.01, 0.02, 0.02, 0.01, 0.93, 0.01]$). Strong central punishment, however, yields the distribution $[0.034, 0.114, 0.204, 0.352, 0.102, 0.193]$, i.e. ineffective (corrupted) pool-punishers dominate, followed by loners and corruptors.

Strong centralized punishment allows corruptors to exploit pool-punishers in two ways: pool-punishers contribute to a public good, at the same time funding a flawed institution that corruptors use to their advantage. Weak centralized punishment, on the other hand, provides an opportunity for peer-punishers to counteract both corruptors and defectors.

Lastly, we introduce $H$ hybrid punishers. In addition to contributing $c$ to the public good, individuals using this strategy pay both $G$ to punish defectors directly and $G$ to the punishment pool, and as such they are not punished by the central authority. Hybrid individuals can be thought of as upstanding citizens that pay their taxes but also engage in forms of 'legitimate' peer sanctioning.

Figure 2 shows that, unlike peer punishment alone, this hybrid strategy dominates the population when centralized punishment is severe. This occurs even though the hybrid strategy pays a higher average cost compared with pool-punishers. Setting $M = 100$, the long-run distribution of strategies in the $(X, Y, Z, V, W, C, H)$ subpopulation is $[0.001, 0.004, 0.008, 0.013, 0.004, 0.007, 0.96]$. 

Figure 2. Stationary distributions of strategies as a function of institutional punishment severity (parameter $B$). In (a), the corruptor strategy is included in the set of available strategies and we observe the adverse effect of institutional punishment. The greater $B$, the greater the percentage of corruption. A clear phase transition happens when $B > (M - 1/N - 1)G$, when the expected punishment exerted by a single pool-punisher (in a sample of $N$) exceeds the punishment cost for the pool-punisher, $G$. This allows a pool-punisher to severely suppress peer-punishers, which in turn allows corruptors, defectors and loners to grow in the population.

In (b), both the corruptor and the hybrid strategies are included. As a result, increasing $B$ no longer backfires, and the same level of cooperation is maintained. The hybrid strategy becomes dominant for $B > (M - 1/N - 1)G$, when the expected punishment exerted by a single peer-punisher (in a sample of $N$) exceeds the punishment cost for the pool-punisher, $G$. 

\[ B = \frac{M - 1}{N - 1} G \]
As a consequence, a high level of cooperation is maintained across all levels of centralized punishment. The dominance of the hybrid strategy is robust against different parameter values (including \( B, y \) and \( K \), as we show in the electronic supplementary information).

Of course, one might wonder why individuals would create a second-order punishment institution in the first place, as figure 2b also shows that second-order punishment does not increase the overall level of cooperation; nor does it make cooperation significantly more stable than peer punishment alone. Our relatively simple model is unlikely to fully answer this very general question, as we have left out many features that could cause centralized institutions to remain advantageous. For example, these institutions might aggregate views on who should be punished; and this aggregation could cause perceptual errors (which are not in our model) to cancel out [35].

It is also possible that institutions may further evolve to deal with this remaining instability. Analytical results in the electronic supplementary information show that when second-order punishment is strong, hybrid punishers are only destabilized by neutral-drift towards pool-punishers (who then allow corruptors and defectors to emerge). Institutions may therefore want to screen and punish pure pool-punishers; and it is interesting that many justice systems have evolved rules that fine people who merely pay their taxes but do not register for various forms of hybrid punishment, for example jury duty.

Importantly, however, we have shown that simply adding the risk of corruption can help to explain why centralized and decentralized forms of punishment frequently coexist. No additional appeal to civic norms or civic culture is needed. Which is not to say that these things do not exist or that they do not further promote citizen participation in policing the commons. Rather, our model shows that independent of other virtues, peer-punishment strategies can have a fitness advantage over pool punishment alone. In the face of corruption, peer and hybrid punishment strategies better promote cooperation because they are competitive. If one punisher fails to punish a corrupt individual, another might step in; and this result may help to explain why polities who want to control corruption and promote cooperation often become more tolerant to various forms of decentralized sanctioning [36].

Funding statement. S.A. is funded by the British University in Dubai. M.C. is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence programme.

References

democracy work: civic traditions in modern Italy.
30. Verba S, Schlozman KL, Brady HE. 2005 Voice and
equality: civic voluntarism in American politics.
Cambridge, MA: Harvard University Press.
31. Fowler JH. 2005 Altruistic punishment and
USA 102, 7047 – 7049. (doi:10.1073/pnas.
0500938102)
32. Hauert C, De Monte S, Hofbauer J, Sigmund K. 2002
Volunteering as Red Queen mechanism for
cooperation in public goods games. Science 296,
1129 – 1132. (doi:10.1126/science.1070582)
33. Hauertwz C, De Montewy S, Hofbauervw J,
Sigmund K. 2002 Replicator dynamics for
34. Semmann D, Krambeck HJ, Milinski M. 2003
Volunteering leads to rock–paper–scissors
dynamics in a public goods game. Nature 425,
390 – 393. (doi:10.1038/nature01986)
35. McLennan A. 1998 Consequences of the Condorcet
jury theorem for beneficial information aggregation
by rational agents. Am. Polit. Sci. Rev. 92,
poor dictators allow freer media: a theory and
evidence from panel data. Am. Polit. Sci. Rev. 103,
645. (doi:10.1017/S0003055409990219)