Hugh-Jones, David; Zultan, Roi

Working Paper

Brothers in Arms: Cooperation in Defence

Jena Economic Research Papers, No. 2010,064

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Hugh-Jones, David; Zultan, Roi (2010) : Brothers in Arms: Cooperation in Defence, Jena Economic Research Papers, No. 2010,064, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/56905

JENA ECONOMIC RESEARCH PAPERS

2010 – 064

Brothers in Arms: Cooperation in Defence

by

David Hugh-Jones
Ro´i Zultan

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Jena Economic Research Papers 2010 - 064

Brothers in Arms: Cooperation in Defence

David Hugh-Jones and Ro'i Zultan

September 24, 2010

Abstract

In experiments, people behave more cooperatively when they are aware of an external threat, while in the field, we observe surprisingly high levels of cooperation and altruism within groups in conflict situations such as civil wars. We provide an explanation for these phenomena. We introduce a model in which different groups vary in their willingness to help each other against external attackers. Attackers infer the cooperativeness of a group from its members' behaviour under attack, and may be deterred by a group which bands together against an initial attack. Then, even self-interested individuals may behave cooperatively when threatened, so as to mimic more cooperative groups. By doing so, they drive away attackers and increase their own future security. We argue that a group's reputation is a public good with a natural weakest-link structure. We test the implications of our model in a laboratory experiment.

Keywords: cooperation, conflict, defence, signaling

JEL Classification: C73, C92, D74

1 Introduction

. Many people behave altruistically towards strangers, even when they are in anonymous, short-term or one-shot interactions. This presents a puzzle for social and biological scientists, because altruism – helping another at a cost to oneself – ought to be selected against. While various explanations have been proposed, in this paper, we focus on one particular aspect of the puzzle: that altruistic behavior increases in the face of an external threat. For example, Bornstein and Ben-Yossef (1994) showed in a laboratory experiment that group members' contributions to a public good increased when they were made aware of a rival group, even though this group could not affect their payoffs in any way. Hargreaves-Heap and Varoufakis (2002) split participants into two groups and created a situation in which one group suffered discrimination; subsequently, pairs of members of that group cooperated more often in a Prisoner's Dilemma than pairs from the other group.

Related phenomena exist in the field. Defence is a canonical example of a "public good", whose provision benefits not only the providers, but also free-riders who contribute nothing. Economic theory predicts that defence will be underprovided unless the state enforces contributions. However, in many civil wars, people fight for their group against other groups, in the absence of state coercion. Some people may be coerced into participation by other group members. While we do not underestimate this aspect of the phenomenon, we do not believe that it can be a general explanation for everybody's participation, and in many historical episodes it seems unlikely to have played a large role. For instance, the risks from taking an active part in the French Resistance, or the Provisional IRA during the Troubles, were surely much higher than any risk one's own side might impose for not taking

part. On a more everyday level, humans appear to become more supportive of their in-group when they face an external threat. To give some examples: there was a large increase in blood donations after the September 11 attacks (Glynn et al., 2003). There is a well-known "rally round the flag" effect in which expressed support for political incumbents increases after a military or terrorist attack (Baker and Oneal, 2001). Lastly, in a recent study of the Israeli small claims court, show that Jewish judges were more likely to find in favour of a Jewish plaintiff against an Arab defendant on the day after a terrorist attack (while Arab judges were more likely to find in favour of the Arab defendant).

Social psychologists have long been aware of this phenomenon, and have argued that "war with outsiders... makes peace inside" (Sumner, 1906; Campbell, 1965). Social identity theorists explain that individuals' sense of group identity is increased by perceived threats to the group (Stephan and Stephan, 2000). While these theories offer insight, they give only a proximate, not an ultimate explanation. We still do not know how humans might have evolved a psychological mechanism that responds to external threats by increasing group identity (and hence encouraging altruistic behaviour, with associated costs to one's own fitness). Indeed, the same question arises in biology, since some species seem to help unrelated conspecifics against predators: examples include defensive rings, mobbing of predators and alarm calls (Edmunds, 1974).

A common explanation for altruistic behaviour is that individuals are in long-term relationships, in which present help will be reciprocated in the future, while not helping may end the relationship and lose the benefits of future cooperation (Trivers, 1971; Rubinstein, 1979; Axelrod and Hamilton, 1981). However, we

believe this is not the whole explanation for cooperation in defence, for various reasons. First, as mentioned above, we observe such cooperation in short-term laboratory interactions. Second, in real-world episodes of group defence, the immediate cost of cooperation is often an immediate risk of injury or death, which is likely to outweigh the future benefits of any reciprocal behaviour by those helped. Lastly, we observe defensive helping in animals, but very few examples of reciprocal altruism have been found in the biological world.

In this paper we suggest a mechanism that may allow helping behavior to evolve among individuals who belong to the same group nominally, but have no other meaningful interaction, when those individuals come under attack by, for example, a rival ethnic group, or a biological predator. The logic is that of signalling. The argument runs as follows:

1. Groups vary in their willingness to cooperate against attackers. This can be for a variety of reasons. In the social world, some groups may indeed be engaged in long-term cooperative relationships, and may help each other both because they expect helping to be reciprocated in future, and because they will suffer from the loss of a partner (Eshel and Shaked, 2001). Hunter-gatherer communities, who acquire food by cooperative game hunting, are an example. Other groups, such as small farmers, may be composed of individually self-sufficient households without a direct incentive to cooperate. Biologically, some groups may be composed of closely related kin, with high mutual altruism, while others are made up of unrelated individuals.

2. Attackers are opportunistic: they attack in order to acquire group members' resources (or, in the case of biological predators, for food). They are

therefore more willing to attack group members if they expect low levels of cooperation in defence. Conversely, if they expect a strong defence from a group, they may prefer to engage in an alternative, less risky activity, or to find a different group to attack.[1]

3. Because of the previous point, attackers have an interest in finding out the type of group they are facing. However, they are not always able to observe a group's level of cooperativeness from the outside. Instead, they will find it optimal to make one or more initial attacks, in order to gauge the cooperativeness of a particular group. They can then decide whether to continue attacking or to break off.

4. As a result, all group members have an interest in appearing cooperative, at least during the initial stages of an attack. By doing so, they may deter the attacker, and prevent future attacks which would eventually fall on themselves. In game-theoretic terms, less cooperative groups have an incentive to *pool* with more cooperative groups.

We model this logic in a simplified setup. Some groups (henceforth *hunters*) participate in social interaction, and will therefore help their fellows who come under attack, whereas other groups (henceforth *gatherers*) have weak intragroup connections and therefore are not motivated to help their peers. An attacker makes one or more attacks on a group; during each attack, the (randomly selected) target individual may be helped by another randomly selected individual, at a cost to the helper which the helper privately observes. After each attack, the attacker may

[1] We treat attackers as single self-interested agents, thus abstracting away from two-sided group conflicts. This would be an interesting extension to the theory.

break off and attack a new group.

When the maximum number of rounds is large enough, this model has a unique equilibrium that survives a natural refinement. The equilibrium has the following characteristics. First, for any group size, so long as individuals are patient enough, helping behaviour can be sustained, even for arbitrarily large costs. Second, the costs borne by defenders in mutual aid may be larger than the actual benefit provided to the helped individual. These two results come from the same fact: the motivation to help is provided not by the benefit to the target, but by the deterrence effect of driving off an attacker; so, the relevant benefit is the difference between being attacked and being left alone, summed over all future rounds. We believe that the second result in particular may help to explain how, in human conflicts, seemingly trivial incidents such as insults of a group member may lead to disproportionate responses.[2] Third, it is irrelevant what proportion of the groups are actually hunters: this can be arbitrarily small. Fourth, cooperation among the gatherers becomes less likely in the face of repeated attacks.[3]

Lastly, cooperation is subject to sudden collapses: if a single individual does not help, then everyone else stops helping. This is closely tied to the signalling logic of the game. An individual who doesn't help provides a certain signal to the attacker that he is facing gatherers, not hunters. Afterwards the attacker can no longer be deterred, and this removes the incentive for other group members to help. Again, there is a real-world analogue in conflict behaviour. In turn, the fact that not helping causes others not to help in future strengthens the motivation to

[2]Many examples of this can be found in Horowitz (2001). Stephan and Stephan (2000) discuss "symbolic threats" from a social psychological point of view.

[3]In equilibrium, the attacker moves on at once after observing a single episode of helping, so this statement holds for off-path behaviour.

help. One might expect that group reputation, like defense itself, would be a public good and therefore be underprovided. However, in our model the reputation good is "produced" using a weakest-link technology, since a single uncooperative action undermines the defenders' reputation. Also, if the signalling logic of cooperation in defence increases the sensitivity of individuals' to each others' behaviour, then this ought to be observable in laboratory experiments. This prediction goes beyond the standard social psychology claim that group identity increases in response to threat. We can therefore use it to test our theory.

Some of the field evidence discussed above, such as ethnocentrism in court judgments, is hard to rationalize as optimal self-interested behaviour. However, the theory can be viewed either as a direct game-theoretic rationalization of helping behaviour within conflict, or, more indirectly, as describing a possible logic behind the evolution of psychological dispositions to cooperate when threatened by attack. That is, these dispositions may have evolved in strategic situations like those of the model, in which small groups faced opportunist external enemies and needed to deter them. If so, these evolved dispositions might still work the same way in larger and more specialized modern societies (cf. Cosmides and Tooby, 1992).[4] In our experiment, we examine how sensitive cooperation is to specific strategic aspects of the situation.

In an extension to the model (section 5), we extend our logic to public goods games which are played among defenders before the attacker decides to attack. Thus, we can rationalize the evidence for increased in-group cooperation in the

[4]Note that, since the model generates predictions about the dynamics of cooperation under threat, beyond existing work in social psychology, we cannot be accused of merely explaining existing data with a "just-so story".

face of external threats, described above.

We believe our paper may be of interest to the following groups.

First, students of conflict, including economists and political scientists, who seek microfoundations for voluntary participation in collective violence. Ideas from game theory have long been found useful to political scientists; notable examples include the Security Dilemma (Jervis, 1978; Posen, 1993) and the divide-the-dollar model of war (Fearon, 1995). But there has been a fundamental block to the wider acceptance of these explanations: collective conflicts seem to violate individual rationality, because they involve individuals voluntarily cooperating (perhaps at great risk) to gain a collective benefit.[5] We offer a possible explanation.

Second, biologists seeking explanations for the evolution of cooperation in defence. Signaling explanations of altruism are well-known in theoretical biology (Zahavi, 1975; Gintis, Smith and Bowles, 2001). In these models helping behavior is a costly signal of individual quality, which benefits the individual helper by (e.g.) making him or her a more attractive partner for reproduction. By contrast, in ours, helping behaviour signals a fact about the group, and benefits the whole group.

Lastly, economic theorists interested in reputation-building may be interested in our model. Previous work has examined reputation-building in repeated games, either with one patient player against an infinite set of short-run players, or with two patient players. Here, we introduce group reputation – since types are perfectly correlated within groups – in a dynamic setup (cf. Healy 2007). The

[5]For example: "... any act by an individual against a large group,... is inherently irrational in the Olsonian sense" (Petersen, 2002); "...pursuit of individual self-interest does not explain torture, murder or risking one's own life in battle" (Kaufman, 2001).

modelling technique in the literature is to assume that a (small) proportion of the reputation-building players is a "Stackelberg type" who always plays the action that gives him the long-term best response, assuming the other players best respond. Similarly, our "hunter types" play so as to maximize the welfare of their group.

2 Model

The "defenders" are a large population of groups of size N.

An attacker makes one or more attacks on a randomly chosen member (the "target") of a randomly chosen group. Another randomly chosen member of the same group (the "supporter") may assist the target at a cost c to its own fitness. The attack costs the defender A and gives the attacker a benefit of A if the helper does not help, and costs the defender/benefits the attacker $a < A$ if the helper helps. We normalize defender per-round welfare so that it is 1, or 0 after a successful attack.[6]

A proportion π of the groups are "hunters", meaning that their members always help the target; the rest are "gatherers". Several different interpretations are possible. Hunters may be altruistic towards one another, perhaps because they are genetically related, while gatherers are purely self-interested. Alternatively, hunters may be in long-term relationships, beyond the scope of the attack episode, and able to enforce cooperation by conditioning their future behaviour on play during the attack episode, whereas gatherers do not expect to interact after the attack

[6] We could assume that defenders are killed and removed from the game. This would strengthen our results by providing another reason for unrelated defenders to help: after a successful attack, the group size shrinks and the helper is more likely to be targeted in future.

episode.

After every attack, the attacker may stay, or may costlessly move to a different group, from a large population of groups. (So large that the chance of returning to the same group later is effectively 0.) However the attacker may make no more than T attacks on any one group.[7] Defenders and attackers share a discount rate δ. There are N defenders in each group. The individual cost of helping c is random and drawn independently in each round from $\mathbf{C} \subset \mathbb{R}^+$, with cdf $\Phi(C) = Pr(c \leq C)$. We assume this is continuous. Only the supporter observes c in each round. We assume that

$$\Phi(\bar{C}) < 1, \text{ where } \bar{C} = \frac{\delta}{1-\delta}\frac{A}{N}. \tag{1}$$

The defenders and the attacker observe the history of attacks within a given group, and whether the target was helped in each case.

3 Equilibrium analysis[8]

The set of histories of length t is $\mathscr{H}^t = \{0,1\}^t$, where 1 indicates that the defender was helped, with typical element h_t. (Write $\mathscr{H}^0 = \emptyset$.) The set of all histories is $\mathscr{H} = \bigcup_{t=0}^{T} \mathscr{H}^t$. A strategy for the attacker is $\zeta : \mathscr{H} \to [0,1]$, giving the probability of playing *stay* after each history. (We will often write $\zeta(h) \in \{stay, move\}$ for clarity: i.e., define $stay = 1$ and $move = 0$.) A pure strategy for a (gatherer)

[7] We use finite repetitions so as to avoid folk-theorem style results where there are multiple equilibria even if the attacker does not condition on defender behaviour: we want to focus on the stark case where repeated play among defenders alone could not sustain cooperation. This also enables us to find a unique equilibrium.

[8] Since the hunters are non-strategic actors, the following analysis deals strictly with gatherers.

defender is $\sigma : \mathscr{H} \times \mathbf{C} \to \{0,1\}$, giving the probability of helping.[9] (Hunters are not strategic actors: they are assumed to always help.) The attacker's subjective probability that he is facing a group of hunters is $\mu : \mathscr{H} \to [0,1]$.

Define p_t as the t-length history of 1s, i.e. the t-length history in which supporters always helped, and let $p_0 = \emptyset$. Let $\mathscr{P} = \{p_0, p_1, p_2, ...\}$. We call these "histories of (perfect) helping". We look for the following equilibrium strategies.

- If the defender has always been helped in the past, the attacker moves to a different group. Otherwise, the attacker attacks the same group forever. Thus $\zeta(h) = move$ if $h \in \mathscr{P}$ and $\zeta(h) = stay$ otherwise.
- Defenders help at round t (after a history h_{t-1}) if and only if (1) all previous defenders have helped (2) c is less than a finite cutpoint C_t. Formally, $\sigma(h_{t-1}, c) = 1$ if $h_t \in \mathscr{P}$ and $c \leq C_t$; $\sigma(h_{t-1}, c) = 0$ otherwise.

Notice in particular that the attacker moves after observing a single episode of helping. Because of this, histories $p_2, p_3, ...$ are off the equilibrium path. In order to ensure reasonable attacker beliefs at these histories, we use the sequential equilibrium concept.

Proposition 1. *For T high enough, the game has a Sequential Equilibrium of the above form (along with appropriate beliefs).*

The remainder of this section gives the proof.

[9] Technically a defender could condition behaviour on his own costs of helping in previous rounds when he was a supporter. Allowing this would not affect our results.

3.1 Supporter behaviour

Given the attacker's strategy, and other defenders' strategies, if at round t $h_t \notin \mathscr{P}$ then a supporter's play does not affect future events in the game (future supporters will never help, and the attacker will always stay). Since $c > 0$ it is never optimal to help.

If at round t, $h_t \in \mathscr{P}$, then the supporter's behaviour determines future play. Helping will cause the attacker to move and not helping will cause the attacker to stay and all future supporters not to help. Thus helping is optimal if

$$1 - c + \sum_{s=1}^{T-t} \delta^s \geq 1 + \sum_{s=1}^{T-t} \delta^s (1 - \frac{A}{N})$$

equivalently

$$c \leq C_t = \frac{\delta - \delta^{T-t+1}}{1-\delta} \frac{A}{N}. \tag{2}$$

C_t is decreasing in t, and in particular, $C_T = 0$. Also, since $C_t < \frac{\delta}{1-\delta} \frac{A}{N} = \bar{C}$, there is always positive probability that the supporter does not help.

3.2 Attacker behaviour

Given these cutpoints, we can calculate the attacker's beliefs. The initial belief $\mu(\emptyset) = \pi$. Since only gatherers fail to help, $\mu(h_t) = 0$ unless $h_t \in \mathscr{P}$.[10]

Write $V(h_t)$ for the attacker's equilibrium value after a history h_t, and $V = V(\emptyset)$. Also, write

$$V_S(h_t)$$

[10]This is shown for beliefs off the path of play in Lemma 5, where the sequential equilibrium refinement is used.

for the attacker's value after h_t if he stays, and subsequently plays his equilibrium strategy.

Equilibrium strategies give

$$V(h_t) = V_S(h_t) = \sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V, \text{ if } h_t \notin \mathscr{P}. \tag{3}$$

In other words, after observing any non-helping, the attacker stays and receives A per round until the number of rounds is up.

Otherwise,$V(h_t) = V$ since the attacker moves (or has just arrived). To show that these are a best response, we can apply the One-Shot Deviation Principle: to check if a strategy is a best response, we need only compare it against deviations involving a single action at one information set.[11] Thus, we need to show that

$$V(h_t) \geq V \text{ if } h_t \notin \mathscr{P},$$

so that after observing a failure to help, it is optimal for the attacker to stay. This is true by (3) and the fact that $V \leq \sum_{s=0}^{\infty} \delta^s A$ given that the attacker's maximum per-round payoff is A. We also need to show that

$$V \geq V_S(h_t) \text{ if } h_t \in \mathscr{P} \tag{4}$$

so that after observing helping it is optimal for the attacker to move rather than to stay. The right hand side here is the counterfactual value from staying for a further

[11] Hendon, Jacobsen and Sloth (1996) prove the principle for Sequential and Perfect Bayesian Equilibrium.

attack. This can be calculated as

$$V_S(h_t) = \mu(h_t)[a+\delta V] + (1-\mu(h_t))\{\Phi(C_{t+1})[a+\delta V] + (1-\Phi(C_{t+1}))[A+\delta V((h_t,0))]\} \text{ if } h_t \in \mathscr{P}.$$

Here, the first term is the value if one is facing hunters: the supporter helps, so the attacker receives a and then moves at once. Similarly, if the attacker is facing gatherers but the supporter's cost drawn is lower than the cutpoint, then the supporter helps, the attacker receives a and moves. Finally, if the cost is higher than the cutpoint, the attacker receives A and the game proceeds. In equilibrium, applying (3),

$$V((h_t,0)) = V_S((h_t,0)) = \sum_{s=0}^{T-t-2} \delta^s A + \delta^{T-t-1} V$$

and plugging this into the previous equation gives

$$\begin{aligned} V_S(h_t) &= [\mu(h_t) + (1-\mu(h_t))\Phi(C_{t+1})][a+\delta V]\cdots \\ &\quad +(1-\Phi(C_{t+1}))[\sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V] \text{ if } h_t \in \mathscr{P}. \end{aligned} \tag{5}$$

We now show that for T high enough, (4) holds given defender behaviour. First, we show that after enough rounds, it always holds. This is simply because the attacker's subjective probability that he is facing a group of hunters becomes increasingly close to certainty after observing enough rounds of cooperation.

Lemma 1. *For M large enough* (4) *holds for $t > M$.*

Proof. First observe that $V > a + \delta V$ since the attacker's minimum payoff in the first round is a and since the attacker receives A with strictly positive probability

in equilibrium. Therefore, if $\mu(h_t)$ is close enough to 1, (5) will be less than V and (4) will hold.

Next, write $\mu_t \equiv \mu(p_t)$ for short (we will keep using this notation) and use Bayes' rule to write

$$\mu_t = \frac{\pi}{\pi + (1-\pi)\prod_{s=1}^{t}\Phi(C_s)}. \tag{6}$$

Since $\Phi(C_t) < \Phi(\bar{C}) < 1$, μ_t is strictly increasing in t and approaches 1 for large enough t.[12] □

The next part of the argument demonstrates the same for early rounds. This relies on choosing T high enough that C_t is very close to $\bar{C}$. The logic is as follows. Staying and observing a further round of helping has three effects on the attacker. First, it increases his probability that he is facing a hunter group. This encourages him to move to a different group. The other effects are that the end of the T rounds is now closer, and that the defenders' cutpoint decreases somewhat (i.e. $C_{t+1} < C_t$). These effects may encourage the attacker to stay. However, when T is large, they become negligible, since the end of the game is far away and (for that reason) the defenders' cutpoint changes very little. Therefore the first effect dominates.

Lemma 2. *For any M, for T high enough, $V_S(\emptyset) > V_S(p_1) > ... > V_S(p_M)$.*[13]

Combining these Lemmas, along with the fact that $V_S(\emptyset) = V$, we can choose M and T large enough that $V \geq V_S(h_t)$ for $h_t \in \mathscr{P}$, both for $t > M$ and for $t \leq M$ as Equation (4) requires. This completes the proof of Proposition 1.

[12] Technically a little more work is necessary to show that only the beliefs of equation (6) are possible in sequential equilibrium. See Lemma 5 in the Appendix.

[13] Proofs not given in the main text are in the Appendix.

4 Uniqueness

Here we investigate whether there are other equilibria. We continue to write V for the value of the game to the attacker, which is also the attacker's value after choosing *move*. First, we demonstrate that behaviour for $h_t \notin \mathscr{P}$ is always the same as in the equilibrium above. The argument is essentially by backward induction: after the attacker has become certain he is facing a group of gatherers, then he cannot be driven off by any further helping, and then cooperation cannot be preserved among the defenders since the game has finite periods.

Lemma 3. *Suppose* $\mu(h_t) = 0$. *Then in any equilibrium,* $\zeta(h_t) =$ *stay and* $\sigma(h_t, c) = 0$ *for all c.*

Sequential equilibrium ensures that $\mu(h_t) = 0$ for all $h_t \notin \mathscr{P}$,[14] so this Lemma shows that in any equilibrium, when $h_t \notin \mathscr{P}$, $\sigma(h_t, c) = 0$ for all c and $\zeta(h_t) =$ *stay*, just as in the previous section. Therefore, the only source of variation in equilibria must be in different attacker and defender responses to a history of helping p_t.

We now show that for T large enough, there is no equilibrium with $\zeta(p_t) > 0$ for $t \geq 1$. Thus, the equilibrium of the previous section is the unique sequential equilibrium. [15]

The proof works as follows. First, we observe that for t large enough, $\zeta(p_t) =$ *move* since it becomes increasingly certain that the defenders are hunters. Next,

[14] See Lemma 5 in the Appendix.

[15] There may be Weak Perfect Bayesian equilibria with $\zeta(p_1) = 0$ (i.e. *move*), $\zeta(p_t) > 0$ for some $t > 1$, in which case, p_t is never reached in equilibrium. However, all Weak Perfect Bayesian equilibria have $\zeta(p_1) =$ *move*.

we show that when there are enough rounds, the defenders' cutpoint is higher at the end of a set of periods for which the attacker stays with positive probability even after observing helping, than at the beginning of these periods. The logic is that at the end, one's own action decides whether the attacker will leave or not. At the beginning, on the other hand, the attacker will stay until some future round and will then only leave if all other supporters have also helped. Thus, the incentive to help is greater in the later round. On the other hand, the future history of play which one can affect may be shorter in the later round; but when T is large enough, this makes little difference.

We then examine the attacker's value at round F, the last round in which $\zeta(p_F) > 0$, and at the last earlier period $L-1$ at which $\zeta(p_{L-1}) = 0$ (or if there is none such, at the beginning of the game). At F the attacker's belief that he is facing a group of hunters is strictly higher, and (as we showed) the cutpoint of gatherers is also higher. Combining these facts reveals that, since the attacker is more likely to observe a further round of defense $V_S(p_{L-1}) > V_S(p_F)$. By our assumption that at $L-1$, moving is optimal,$V \geq V_S(p_{L-1})$. Thus, we arrive at $V > V(p_F)$, which contradicts the assumption that staying is optimal at p_F.

Proposition 2. *For T large enough, $\zeta(p_t) =$ move for all $t \geq 1$.*

So far we have used a tool of "rationalist" game theory. Given our applications to biology, and our argument that signalling logic affected the evolution of human dispositions to cooperate, it is interesting to ask whether the equilibrium of Section 2 is evolutionarily stable. Technically, it is not an Evolutionarily Stable Strategy, since both defenders and attackers may play differently at histories which are not on the equilibrium path (for example, p_t for $t \geq 2$), without affecting their

welfare. However, for T large enough, all Weak Perfect Bayesian equilibria satisfy $\zeta((1)) = move$ (and C_1 as defined in (2), and $\zeta(h) = stay$ and $\sigma(h,c) = 0, \forall c$, for $h \notin \mathscr{P}$). It would therefore be surprising if the equilibrium outcome given by these actions were not evolutionarily stable.

5 Cooperation before conflict

In the introduction we mentioned the evidence that cooperative and helping behaviour seems to increase when there is an attack, or the threat of an attack, from the outside. We can extend the model to give a natural explanation for this. The setup is kept as simple as possible to focus on the intuition.

Suppose now that the attacker must commit before the game to attacking for all T periods, or moving. This resembles an irrevocable decision to launch a war. In the period before making his choice, the attacker observes K randomly selected group members playing a Prisoner's Dilemma. Player i's cooperation gives $R \in (1/K, 1)$ to each member of the group, at a cost of q to the player where q is drawn from a distribution with pdf $\Psi(\cdot)$, supported on $(R, 1)$. The value of q is common knowledge among defenders but not the attacker. As before, hunters always cooperate. After observing play in the Prisoner's Dilemma, the attacker either attacks, or does not, earning a payoff of P. This could be the expected payoff from attacking a different group, or the payoff from some other activity.

In the attacks, hunters always help and gatherers never help, since the attacker

cannot be deterred.[16] We assume

$$\sum_{t=1}^{T} \delta^t \frac{a}{N} < P < \sum_{t=1}^{T} \delta^t \frac{A}{N}.$$

The expected loss to each defender from facing an attack is:

$$\sum_{t=1}^{T} \delta^t \frac{A}{N}.$$

There is always an equilibrium in which gatherers do not cooperate. However, there may also be cooperation in equilibrium, for the same signalling reason as before. We seek an equilibrium in which all gatherers cooperate if $q \leq \bar{q}$.

It must be the case that such cooperation (and only such cooperation) deters the attacker. The attacker's belief after observing full cooperation is

$$\mu = \frac{\pi}{\pi + (1-\pi)\Psi(\bar{q})} \tag{7}$$

and he is deterred if

$$\mu \sum_{t=1}^{T} \delta^t \frac{a}{N} + (1-\mu) \sum_{t=1}^{T} \delta^t \frac{A}{N} \leq P. \tag{8}$$

If he observes any non-cooperation he learns for sure that the defenders are gatherer types, and attacks (since $\sum_{t=1}^{T} \delta^t \frac{A}{N} > P$).

Since μ in (7) is decreasing in $\bar{q}$, (8) provides an upper limit for $\bar{q}$. Above this

[16]The Prisoner's Dilemma itself may be the basis for the differentiation between group types. For example, hunters can be engaging in the game repeatedly with the same partners, whereas the gatherers often reconstruct new groups with stranger members. The attacker observes only one period of the repeated game, and therefore cannot distinguish between partner and stranger groups.

upper limit, cooperation is not convincing enough since too many gatherer types are doing it. Call this the "attacker deterrence constraint".

If the attacker is deterred by full cooperation, and $q \leq \bar{q}$ so that other defenders will cooperate, then it is optimal for each defender to join in cooperating if

$$R - q \geq -\sum_{t=1}^{T} \delta^t \frac{A}{N},$$

equivalently if

$$q \leq R + \sum_{t=1}^{T} \delta^t \frac{A}{N}.$$

This provides another upper limit on $\bar{q}$. Call it the "reward constraint", since it requires that the reward from cooperation be large enough to justify the cost. Of course, $\bar{q}$ may be lower than these, since no defender will cooperate if, for a given value of q, he or she expects the others not to cooperate. To sum up, there is a set of equilibria in which gatherer defenders cooperate for $q \leq \bar{q}$ where

$$0 \leq \bar{q} \leq \min\{R + \sum_{t=1}^{T} \delta^t \frac{A}{N}, \hat{q}\}$$

where

$$\hat{q} \equiv \Psi^{-1}\left(\frac{\pi}{1-\pi}\left(\frac{\sum_{t=1}^{T} \delta^t \frac{A-a}{N}}{\sum_{t=1}^{T} \delta^t \frac{A}{N} - P}\right)\right)$$

is the solution to (7) and (8).

Examining the upper bound for $\bar{q}$ reveals the following. (1) If only the attacker's deterrence constraint is binding, so that the upper bound is given by $\hat{q}$, then it is weakly increasing in P and π. An increase the value of the outside option, or in the probability the attacker puts on the defenders being hunters, will make him easier

to deter. Also, in this case the upper bound is decreasing in A[17] and a: a greater benefit for the attacker from finding either kind of group makes him harder to deter. Finally, the upper bound increases if Ψ increases (in the sense of first order stochastic dominance): when average costs get higher, then cooperation up to a higher cost level will still persuade the attacker that he is facing a hunter group. (2) If only the reward constraint is binding then the upper bound is increasing in R and A: cooperation is sustainable at higher levels when it is more efficient in itself, and when the cost of an attack is high.

It is clear that this logic could be extended to many different game forms, including episodes of pairwise cooperation or altruism – any behaviour that correlates with the desire to cooperate in an actual attack.

6 When history is unobserved

We return to the framework of Section 2, in order to make a slight modification. Some readers may be concerned that our result is driven by the history-dependent behaviour of other defenders. Since future supporters will cease to help if the current supporter does not help, perhaps this is just a Folk-theorem like result albeit for finite repetitions. To show this is not so, we now assume that defenders cannot condition on others' behaviour. Instead, a gatherer strategy is $\sigma : \{1,...,T\} \times \mathbf{C} \to \{0,1\}$, where $\sigma(t,c)$ gives the probability of helping in each round t, given a helping cost of c.

We look for an analogue of the earlier equilibrium, in which the attacker is in-

[17] To show this, differentiate $\hat{q}$, recalling that $P > \sum_{t=1}^{T} \delta^t \frac{a}{N}$.

stantly deterred by a single episode of helping on the equilibrium path.

Proposition 3. *If and only if* $\Phi(\frac{\delta}{1-\delta}\frac{A}{N}) < \frac{\sqrt{\pi}-\pi}{1-\pi}$*, then for large enough* T *there is an equilibrium of the following form:*

$\zeta(h) =$ *move if and only if* $h \in \mathscr{P}$.

Gatherer defenders help during the first attacks if and only if c *is less than* $C_1 = \sum_{t=1}^{T-1} \delta^T \frac{A}{N}$*. In subsequent attacks they never help.*

The expression $\frac{\sqrt{\pi}-\pi}{1-\pi}$ is increasing in π and approaches 0 as $\pi \to 0$. Thus, our conclusions are modified somewhat when defenders cannot condition on each others' behavior. Our equilibrium only exists when the proportion of hunters is non-negligible, compared to the probability of low costs.

[other extensions

... with defenders being "killed" (i.e. group size shrinking) after a successful attack...

... if related defenders do not always defend ...

... if the attacker pays a cost to move between groups ... more generally, if there is an exogenous outside option from leaving.]

7 Conclusion

There has been a recent surge of interest in the economics of conflict. Yet economic models of conflict typically lack microfoundations. This paper provides one: actors may cooperate against an outside attacker in order to drive him/her/it

off by appearing like a highly cooperative group. Resulting cooperation levels decrease in group size, but they can be arbitrarily high if the time horizon of the attack is long enough and defenders are patient enough. Also, they do not necessarily depend on the proportion of truly cooperative groups in the population as a whole.

We see scope for further work in the following areas. First, can the uniqueness result be generalized to a wider class of games with group reputation? Second, extending the model to have two groups in conflict, rather than one group and one unified attacker, would help us to understand the logic of civil wars and ethnic conflict. Lastly, in our theory, defensive cooperation is due to group members' expectations of further attacks. In the model, groups are exogenously given. However, a group might also be defined by the attacker's (perhaps arbitrary) choice of targets. This would provide a model of violence and the social construction of identity, as argued for in Fearon and Laitin (2003).

References

Axelrod, R. and W. D. Hamilton. 1981. "The evolution of cooperation." *Science* 211(4489):1390.

Baker, W. D. and J. R. Oneal. 2001. "Patriotism or Opinion Leadership?: The Nature and Origins of the "Rally Round the Flag" Effect." *Journal of Conflict Resolution* 45(5):661–687.

Bornstein, G. and M. Ben-Yossef. 1994. "Cooperation in inter-group and single-group social dilemmas." *Journal of Experimental Social Psychology* 30:52–52.

Campbell, D. T. 1965. Ethnocentric and other altruistic motives. In *Nebraska symposium on motivation*. Vol. 13 pp. 283–311.

Cosmides, L. and J. Tooby. 1992. Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture*, ed. L. Cosmides, J. Tooby and J. H Barkow. Vol. 163 Oxford: Oxford University Press p. 228.

Edmunds, M. 1974. *Defence in animals: a survey of anti-predator defences*. Longman Harlow.

Eshel, I. and A. Shaked. 2001. "Partnership." *Journal of Theoretical Biology* 208(4):457–474.

Fearon, J. D. 1995. "Rationalist explanations for war." *International Organization* 49(3):379–414.

Fearon, J. D. and D. D. Laitin. 2003. "Violence and the social construction of ethnic identity." *International Organization* 54(04):845–877.

Gintis, H., E. A. Smith and S. Bowles. 2001. "Costly signaling and cooperation." *Journal of Theoretical Biology* 213(1):103–119.

Glynn, Simone A., Michael P. Busch, George B. Schreiber, Edward L. Murphy, David J. Wright, Yongling Tu and Steven H. Kleinman. 2003. "Effect of a National Disaster on Blood Supply and Safety: The September 11 Experience." *JAMA* 289(17):2246–2253.
URL: *http://jama.ama-assn.org/cgi/content/abstract/289/17/2246*

Hargreaves-Heap, S. and Y. Varoufakis. 2002. "Some experimental evidence on the evolution of discrimination, co-operation and perceptions of fairness." *Economic Journal* pp. 679–703.

Healy, P. J. 2007. "Group reputations, stereotypes, and cooperation in a repeated labor market." *The American Economic Review* pp. 1751–1773.

Hendon, Ebbe, Hans J?rgen Jacobsen and Birgitte Sloth. 1996. "The One-Shot-Deviation Principle for Sequential Rationality." *Games and Economic Behavior* 12(2):274–282.

URL: *http://www.sciencedirect.com/science/article/B6WFW-45V7FNN-7/2/6f9d051e8e6e8968539ce284d3c7ad5d*

Horowitz, D. L. 2001. *The deadly ethnic riot*. University of California Press.

Jervis, R. 1978. "Cooperation under the security dilemma." *World Politics: A Quarterly Journal of International Relations* pp. 167–214.

Kaufman, S. J. 2001. *Modern hatreds: The symbolic politics of ethnic war*. Cornell University Press.

Petersen, R. D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Cambridge University Press.

Posen, B. R. 1993. "The security dilemma and ethnic conflict." *Survival* 35(1):27–47.

Rubinstein, A. 1979. "Equilibrium in supergames with the overtaking criterion." *Journal of Economic Theory* 21(1):1–9.

Stephan, W. G. and C. W. Stephan. 2000. An Integrated Threat Theory of Prejudice. In *Reducing prejudice and discrimination: The Claremont symposium on applied social psychology*. Lawrence Erlbaum p. 23.

Sumner, W. G. 1906. *Folkways: A study of the sociological importance of usages, manners, customs, mores, and morals*. Ginn.

Trivers, R. L. 1971. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology* 46(1):35–57.

Zahavi, A. 1975. "Mate selection–a selection for a handicap." *Journal of theoretical Biology* 53(1):205–214.

Appendix

Proof of Lemma 2

Proof. Rewrite (5) as

$$V_S(p_t) = [\mu_t + (1-\mu_t)\Phi(C_{t+1})][a+\delta V] + [(1-\mu_t)(1-\Phi(C_{t+1}))][\sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V].$$

Now, $\sum_{s=0}^{T-t-1} \delta^s A + \delta^{T-t} V$ is strictly decreasing in t and is greater than $a+\delta V$. Therefore, to show the above is strictly decreasing in t, it will suffice if

$$(1-\mu_t)(1-\Phi(C_{t+1})) \tag{9}$$

is decreasing in t. Rewrite this expression, using the definition of $\mu(h_t)$ in (6), as

$$\left(1-\frac{\pi}{\pi+(1-\pi)\prod_{s=1}^{t}\Phi(C_s)}\right)(1-\Phi(C_{t+1})).$$

Observe from the definition of C_t in (2) that, for any t, $C_t \to \bar{C}$ as $T \to \infty$. Since Φ is continuous, the above expression approaches

$$(1-\bar{\mu}_t)(1-\Phi(\bar{C})) \text{ where } \bar{\mu}_t \equiv \frac{\pi}{\pi+(1-\pi)\Phi(\bar{C})^t} \tag{10}$$

as $T \to \infty$. This expression is strictly decreasing in t, since $\bar{\mu}_t$ is strictly increasing in t. Define $\varepsilon = \min_{t\in\{0,\ldots,M-1\}}(1-\bar{\mu}_{t+1})(1-\Phi(\bar{C})) - (1-\bar{\mu}_t)(1-\Phi(\bar{C}))$ and note that $\varepsilon > 0$. Now, by selecting T large enough, we can ensure that

$$\left|\frac{(1-\pi)\prod_{s=1}^{t}\Phi(C_s)}{\pi+(1-\pi)\prod_{s=1}^{t}\Phi(C_s)}(1-\Phi(C_{t+1})) - (1-\bar{\mu}_t)(1-\Phi(\bar{C}))\right| < \frac{\varepsilon}{2} \text{ for all } t,$$

and this, combined with our definition of ε, ensures that (9) is decreasing. □

Lemma 4. *In any equilibrium, after any history h_t, gatherers do not help with probability of at least $1-\Phi(\bar{C}) > 0$.*

Proof. Gatherers help if

$$1-c+\delta W \geq 1+\delta W'$$

where W and W' are continuation values from helping and not helping respectively. These are bounded below by $\sum_{s=0}^{T-t}\delta^s(1-\frac{A}{N})$ and above by $\sum_{s=0}^{T-t}\delta^s$. The above bound is reached if the attacker leaves; the lower bound holds because the defender can achieve at least this payoff by never helping. The maximum difference between δW and $\delta W'$ is thus $\delta\sum_{s=0}^{T-t-1}\delta^s\frac{A}{N} = \frac{\delta-\delta^{T-t+1}}{1-\delta}\frac{A}{N} < \bar{C}$; so for $c \geq \bar{C}$

the inequality above will not be satisfied. □

Lemma 5. *In any sequential equilibrium, beliefs $\mu(p_t)$ must be as given in equation* (6)*, while $\mu(h_t) = 0$ for $h_t \notin \mathscr{P}$.*

Proof. First, observe that in any equilibrium, defender play $\sigma(p_t, c)$ can be characterized by a (perhaps infinite) cutpoint C_t, because if $\sigma(p_t, c) = 1$ is optimal, then $\sigma(p_t, c')$ must be strictly optimal for $c' < c$. Since p_t may be off the equilibrium path of play, permissible beliefs must be derived by constructing a sequence of equilibria of perturbed games in which (1) defenders' probability of helping at h_t, $\sigma_n(h_t, c)$ is bounded within a subinterval of $(0,1)$, with the interval approaching $[0,1]$ as $n \to \infty$, for all h_t and c; (2) $\sigma_n(h_t, c) \to \sigma(h_t, c)$ as $n \to \infty$ (to avoid complications we assume that this convergence is uniform across all c) and (3) attacker's probability of leaving or staying is similarly bounded between 0 and 1 and converges to 0 or 1 according to$\zeta(h_t) \in \{stay, move\}$. We also assume that gatherer defenders help with probability $1 - \eta_n(h_t, c) \to 1$ as $n \to \infty$. We then apply Bayes' rule to give the attacker's beliefs. For p_t, this results in

$$\mu_n(p_t) = \frac{\pi \prod_{s=1}^{t} \{\int (1 - \eta_n(p_s, c)) d\Phi(c)\}}{\pi \prod_{s=1}^{t} \{\int (1 - \eta_n(p_s, c)) d\Phi(c)\} + (1 - \pi) \prod_{s=1}^{t} \{\int \sigma_n(p_s, c) d\Phi(c)\}}.$$

As $n \to \infty$ we arrive at the limit

$$\mu(p_t) = \frac{\pi}{\pi + (1 - \pi) \prod_{s=1}^{t} \{\int \sigma_n(p_s, c) d\Phi(c)\}}$$

and in the equilibrium of Section 2, since $\sigma_n(p_s, c) \to 1$ for $c \leq C_s$, $\sigma_n(p_s, c) \to 0$

otherwise, this must reduce to

$$\mu(p_t) = \frac{\pi}{\pi + (1-\pi)\prod_{s=1}^{t}\Phi(C_s)}$$

as in (6).

For $h_t \notin \mathscr{P}$, in any equilibrium, write $h_t = (r_1, r_2, ..., r_t)$, with $r_s \in \{0,1\}$ for $s \in \{1,...,t\}$. Bayes' rule gives

$$\mu_n(h_t) = \frac{\pi\prod_{s=1}^{t}\{r_s\int(1-\eta_n(h_s,c))d\Phi(c) + (1-r_s)\int\eta_n(h_s,c)d\Phi(c)\}}{D}$$

with

$$\begin{aligned} D &= \pi\prod_{s=1}^{t}\left\{r_s\int(1-\eta_n(h_s,c))d\Phi(c) + (1-r_s)\int\eta_n(h_s,c)d\Phi(c)\right\} \\ &\quad +(1-\pi)\prod_{s=1}^{t}\left\{r_s\int\sigma_n(p_s,c)d\Phi(c) + (1-r_s)\int(1-\sigma_n(p_s,c))d\Phi(c)\right\}. \end{aligned}$$

Since $r_s = 0$ for at least one s, the numerator of the above expression goes to 0 as $n \to \infty$, and the denominator D remains bounded above 0 since gatherer types sometimes fail to help after any history (Lemma 4). Thus $\mu(h_t) = 0$. □

Lemma 6. *Suppose that $\zeta((h_t,0,h_+)) = \zeta((h_t,1,h_+))$ for all continuation histories h_+ of length 0 or more. Then in any equilibrium, $\sigma(h_t,c) = 0$ for all c.*

Proof. We prove by backwards induction over the T periods. First, in a final period history h_{T-1}, $\sigma(h_{T-1},c) = 0$ for all c, since supporter behaviour cannot affect future play. Next, at $T-2$, $\sigma(h_{T-2},c) = 0$ for all c, since the supporter cannot affect either future supporter play (as we have just shown) or the attacker's

future play (by assumption). Then at $T-3$, $\sigma(h_{T-3},c)=0$ for all c for the same reason, and so on. □

Proof of Lemma 3

Proof. Again, start at the end. Since $\mu(h_{T-1})=0$, the attacker is certain that the defenders are gatherer types, and since $\sigma(h_{T-1},c)=0$ for all c, the attacker will gain his maximum per-round payoff of A next round by staying, giving a continuation value of $A+\delta V > V$ (since there is positive probability of receiving a in the first round, $V < A/(1-\delta)$). Thus $\zeta(h_{T-1}) = stay$ is strictly optimal.

Now consider $\zeta(h_{T-2})$. Since $\mu(h_{T-2})=0$, the attacker's belief will stay at 0 for any continuation history. Thus, $\zeta((h_{T-2},0)) = \zeta((h_{T-2},1)) = stay$ as we have just shown. Therefore, the assumption of Lemma 6 holds for histories of length $T-2$. Applying Lemma 6, we conclude that $\sigma(h_{T-2},c)=0$ for all c. Therefore $\zeta(h_{T-2}) = stay$. For, given that $\sigma(h_{T-2},c) = \sigma((h_{T-2},0),c) = \sigma((h_{T-2},1),c) = 0$ for all c, and that $\mu(h_{T-2})=0$, the continuation value for staying is $A+\delta A+\delta^2 V > V$. We have now proved the conclusion of the Lemma for histories of length $T-2$.

At h_{T-3}, if $\zeta(h_{T-3}) = stay$ then the previous paragraph shows that $\zeta((h_{T-3},h_+)) = stay$ for any positive-length continuation history h_+. Again this allows us to apply Lemma 6 and shows that $\sigma(h_{T-3},c)=0$ for any c, and again this shows that $\zeta(h_{T-3}) = stay$. This plus the previous paragraph proves the conclusion of the Lemma for histories of length $T-3$. Continuing thus, we prove it for histories of any length. □

Lemma 7. *There is some $\bar{t}$ such that in any equilibrium for a game of any length T, $\zeta(p_t) = move$ for all $t \geq \bar{t}$.*

Proof. Applying (6), Lemma 4 shows that in any equilibrium$\mu(p_t)$ is strictly increasing in t, and so approaches 1. Furthermore, in any equilibrium, since the probability of helping is no more than $\Phi(\bar{C})$, $\mu(p_t) \geq \bar{\mu}_t$ as defined in (10). Therefore, the set of beliefs $\mu(p_t)$, defined over all equilibria, approaches 1 *uniformly* as $t \to \infty$: for any $\varepsilon > 0$, there is some $\bar{t}_\varepsilon$ such that $\mu(p_{\bar{t}_\varepsilon}) \geq \bar{\mu}_{\bar{t}_\varepsilon} > 1 - \varepsilon$ in any equilibrium.

Now, the value to the attacker of staying in equilibrium can be written

$$V_S(p_t) = \mu(p_t)[a + \delta V'] + (1 - \mu(p_t))V'' \tag{11}$$

where V' is the continuation value conditional on the defenders being hunter types, and V'' is the value if the defenders are gatherers. Since hunter types always help, the best response when faced with hunters is to leave; therefore $a + \delta V' \leq a + \delta V$. Furthermore,

$$V \geq (\pi + (1-\pi)\Phi(\bar{C}))a + (1-\pi)(1-\Phi(\bar{C}))A + \delta V = a + \delta V + (1-\pi)(1-\Phi(\bar{C}))(A-a),$$

since (1) the probability of gatherers helping is no more than $\Phi(\bar{C})$, and (2) the attacker can achieve at least the payoff on the RHS, by leaving after the first round. Therefore, in any equilibrium, $a + \delta V' \leq V - \varepsilon_2$ where $\varepsilon_2 = (1-\pi)(1 - \Phi(\bar{C}))(A-a)$. Plugging this into (11), and using the fact that V'' is bounded above

by $\sum_{s=0}^{\infty}\delta^s A$, gives for any ε some $\bar{t}_\varepsilon$ such that

$$\begin{aligned} V_S(p_{\bar{t}_\varepsilon}) &\leq (1-\varepsilon)(V-\varepsilon_2)+\varepsilon\sum_{s=0}^{\infty}\delta^s A \\ &\leq V-(1-\varepsilon)\varepsilon_2+\varepsilon\sum_{s=0}^{\infty}\delta^s A \end{aligned}$$

Choosing ε so that the right hand side is strictly less than V for any equilibrium value of V, we can set $\bar{t}=\bar{t}_\varepsilon$. Then, it is sequentially rational to leave after $p_{\bar{t}}$, so $\zeta(p_{\bar{t}})=leave$. □

Proof of Proposition 2

Proof. Suppose false, so that $\zeta(p_t)>0$ for some $t>0$. If $T\geq\bar{t}$, $\zeta(p_t)=0$ (i.e. *leave*) for t high enough, as Lemma 7 shows. So, for T large enough we may take F such that $\zeta(p_F)>0$, but $\zeta(p_{F+1})=0$. Now, define $L=\min\{t\geq 1:\zeta(p_{t'})>0 \text{ for all } t\leq t'\leq F\}$. Observe that if $\zeta(p_t)=0$ for all $t<F$, then $L=F$; if $\zeta(p_t)>0$ for all $t<F$, then $L=1$.

First we show that $C_L<C_{F+1}$. After p_F, the attacker will condition on the next round, staying until T if he observes no helping and leaving otherwise. Thus,

$$C_{F+1}=\frac{\delta-\delta^{T-F}}{1-\delta}\frac{A}{N},$$

just as in (2). Observe that for any T, $F<\bar{t}$, by Lemma 7. Therefore as T becomes large,

$$C_{F+1}\to\bar{C}=\sum_{t=1}^{\infty}\delta^t\frac{A}{N}. \tag{12}$$

Now examine the supporter's problem in round L. The benefit of not helping is

$$1+\sum_{t=L+1}^{T}\delta^{t-L}\left[1-\frac{A}{N}\right]. \tag{13}$$

The benefit of helping is

$$1-c+\sum_{t=L+1}^{F}\delta^{t-L}\left[1-Nohelp_t\frac{A}{N}-Attack_t\left\{\frac{1}{N}\int_0^{C_t}\hat{c}d\Phi(\hat{c})+\frac{1}{N}[\Phi(C_t)a+(1-\Phi(C_t))A]\right\}\right]+\sum_{t=F+1}^{T}\delta \tag{14}$$

where $Nohelp_t$ gives the probability that at least one defender failed to help between rounds $L+1$ and $t-1$, and $Attack_t$ gives the probability that the attacker is still present at time t even though all defenders helped. That is, until round F, the attacker may still be present even after observing helping. If so, the defender bears the expected cost in curly brackets, which includes the expected cost of being a supporter and helping if $c \leq C_t$, and the expected cost of being attacked and perhaps helped. From round $F+1$ onwards, either the attacker has observed perfect helping and left, or $h \notin \mathscr{P}$, the attacker is staying forever and no defenders help.

We can calculate $Attack_t$ as

$$\prod_{s=L+1}^{t-1}\Phi(C_s)\zeta(p_s)$$

which is positive by definition of L, and $Nohelp_t$, recursively, as

$$Nohelp_{t-1}+(1-Nohelp_{t-1})\zeta(p_{t-2})(1-\Phi(C_{t-1}))$$

with $Nohelp_{L+1}=0$ since by assumption the current supporter helped. I.e. even

if every supporter helped up till $t-2$, if the attacker continued to stay then at $t-1$ the supporter may have failed to help. All that matters is that both $Attack_t$ and $Nohelp_t$ are positive, since $\zeta(p_t)$ is positive for $L \leq t \leq F$.

Rearranging (14) and (13), and taking $T \to \infty$, gives

$$C_L \underset{T\to\infty}{\longrightarrow} \sum_{t=L+1}^{F} \delta^{t-L}\left[(1-Nohelp_t)\frac{A}{N} - Attack_t\left\{\frac{1}{N}\int_0^{C_t} \hat{c}d\Phi(\hat{c}) + \frac{1}{N}[\Phi(C_t)a + (1-\Phi(C_t))A]\right\}\right] + \sum_{t=F+1}^{\infty} \delta^{t-L}(1-$$

Comparing this with 12 shows $C_L < C_{F+1}$, since each term of the above sum is less than $\frac{A}{N}$.

Now,

$$V_S(p_{L-1}) = [\mu_{L-1} + (1-\mu_{L-1})\Phi(C_L))](a + \delta V(p_L)) + (1-\mu_{L-1})(1-\Phi(C_L))(A + \delta A + ... + \delta^{T-L}A + \delta$$

where the first term in brackets gives the probability of the supporter helping, and $V(p_L)$ is the value after p_L. Observe that

$$a + \delta V(p_L) < A + \delta A + ... + \delta^{T-L}A + \delta^{T-L+1}V$$

since $V(p_L)$ involves a sequence of no more than$T-L$ attacks which can give no more than A, followed by V, and since $V < A + \delta V$ implies $V < A + \delta A + ... +$

$\delta^{t-1}A + \delta^t V$ for any $t \geq 1$. Therefore we can write

$$\begin{aligned} V_S(p_{L-1}) &> [\mu_F + (1-\mu_F)\Phi(C_{F+1}))](a + \delta V(p_L)) + (1-\mu_F)(1-\Phi(C_{F+1}))(A + \delta A + ... + \delta^{T-L}A + \delta^{T-L+1}V) \\ &\quad \text{(by } \mu_F > \mu_{L-1} \text{ and } C_L < C_{F+1}, \text{ and } a + \delta V(p_L) < A + \delta A + ... + \delta^{T-L}A + \delta^{T-L+1}V) \\ &> [\mu_F + (1-\mu_F)\Phi(C_{F+1}))](a + \delta V) + (1-\mu_F)(1-\Phi(C_{F+1}))(A + \delta A + ... + \delta^{T-F-1}A + \delta^{T-F}V) \\ &\quad \text{(since } V(p_L) \geq V, \text{ as must always hold given that leaving is an option,} \\ &\quad \text{and } V < A + \delta V \Rightarrow \delta^{T-F}V < \delta^{T-F}A + \delta^{T-F+1}A + ... + \delta^{T-L}A + \delta^{T-L+1}V) \\ &= V(p_F). \end{aligned}$$

But since, by definition of L, either $\zeta(p_{L-1}) = 0$, or $V_S(p_{L-1}) = V$ if $L = 1$, it must be that $V \geq V_S(p_{L-1})$. We therefore arrive at $V > V(p_F)$ which contradicts $\zeta(p_F) > 0$. □

Proof of Proposition 3

Proof. First consider defender behaviour. Since $\zeta((1)) = move$, if $t \geq 2$ then the attacker must have observed not helping and will stay forever. Therefore it is not optimal to bear any cost to help. Now suppose that $t = 1$. Helping gives expected welfare of

$$1 - c + \sum_{t=1}^{T-1} \delta^t$$

and not helping gives

$$1 + \sum_{t=1}^{T-1} \delta^t (1 - A/N)$$

giving a cutpoint

$$C_1 = \sum_{t=1}^{T-1} \delta^t \frac{A}{N}.$$

Next consider attacker behavior. Write $p_t = (1,1,..,1)$ for a t-length history of helping, so that $p_t \in \mathscr{P}$. Clearly since only related helpers help in the second and subsequent periods, $v(p_t) = move$ is optimal for $t \geq 2$. The interesting question is $\zeta(p_1)$, the optimal strategy after a single episode of helping. The benefit of attacking is

$$\mu_1(a+\delta V)+(1-\mu_1)(\sum_{t=0}^{T-2}\delta^t A+\delta^{T-1}V)$$

with

$$\mu_1 = \frac{\pi}{\pi+(1-\pi)\Phi(C_1)}$$

while the benefit of moving is

$$V = [\pi+(1-\pi)\Phi(C_1)](a+\delta V)+(1-\pi)(1-\Phi(C_1))(\sum_{t=0}^{T-1}\delta^t A+\delta^T V)$$

We wish to show conditions when the benefit of moving is greater than that of attacking. Taking T to infinity, the relevant inequality becomes

$$\mu_1(a+\delta V)+(1-\mu_1)\sum_{t=0}^{\infty}\delta^t A \leq [\pi+(1-\pi)\Phi(C_1)](a+\delta V)+(1-\pi)(1-\Phi(C_1))\sum_{t=0}^{\infty}\delta^t A.$$

Since $a+\delta V < \sum_{t=0}^{\infty}\delta^t A$, this will hold in the limit whenever $\mu_1 > \pi+(1-\pi)\Phi(C_1)$, equivalently when

$$\frac{\pi}{\pi+(1-\pi)\Phi(C_1)} > \pi+(1-\pi)\Phi(C_1).$$

This results in a quadratic, but we can observe at once that it holds for $\Phi(C_1) \to 0$, does not hold for $\Phi(C_1) \to 1$, and has a single crossover point in terms of $\Phi(C_1)$. Intuitively, when $\Phi(C_1)$ is small enough, the fact that the supporter helped

is strong evidence that the defenders are indeed hunters. Taking $T \to \infty$ gives $C_1 \to \sum_{t=1}^{\infty} \delta^t \frac{A}{N} = \frac{\delta}{1-\delta}\frac{A}{N}$. Solving the quadratic for $\Phi(C_1)$ gives

$$\Phi(C_1) = \frac{\sqrt{\pi} - \pi}{1 - \pi}$$

as the upper bound for $\Phi(C_1)$ for the equilibrium to exist. □