The Road to Hell? Third-Party Intervention to Prevent Atrocities

Andrew H. Kydd  
University of Wisconsin, Madison

Scott Straus  
University of Wisconsin, Madison

Preventing large-scale atrocities has emerged as an important policy goal of the post–Cold War period. However, a debate exists about the effects of creating an international institution to prevent atrocities. Advocates of intervention argue that a credible threat to intervene should deter perpetrators and stop atrocities when deterrence fails. Critics argue that third-party intervention, by strengthening weak minority groups and lowering the cost of war, encourages rebellions and so makes war and atrocities more likely. We develop a model of intervention to analyze this debate. The model shows that the negative effects of intervention highlighted by critics can be mitigated if the third party is relatively neutral and if alternative costs are imposed on decision makers. We conclude that with appropriate institutional design, the net impact of stronger third-party commitments to end atrocities will be to lower the expected level of atrocities.

Since the end of the Cold War, the question of whether to intervene to stop states from committing atrocities, especially mass killing, has become central (Hoffman 1996; Holzgrefe and Keohane 2003; Rotberg 2010; Teson 1997; Weiss 2007; Wheeler 2003). Such international “humanitarian interventions” generally require the coercive deployment of force by an external actor or coalition. Advocates of intervention argue that a stronger antiatrocities regime will often deter states from committing atrocities. If deterrence fails, they argue intervention will save lives and help deter potential future perpetrators. In this perspective, the main problem is generating the “political will” required to intervene (Chalk et al. 2010; Power 2002). Several recent initiatives are designed to institutionalize an antiatrocities policy, including the Responsibility to Protect (R2P) doctrine within the United Nations and the Albright-Cohen report in the United States (Albright and Cohen 2008; Evans 2008). The success of the Libyan intervention, R2P proponents argue, concludes the debate over intervention in their favor (Evans 2011).

Not all observers agree. Critics charge that these policy initiatives are not cost-effective and in some cases are positively harmful (De Waal 2012; Valentino 2011). In particular, some analysts argue that the prospect of intervention encourages weak groups to rebel and even to deliberately provoke state atrocities in order to trigger international intervention on their behalf. Invoking the concept of moral hazard, these scholars argue that greater resolve to intervene in cases of mass atrocity will therefore perversely increase the probability of war and atrocities, in a grim illustration of the adage that the road to hell is paved with good intentions (Crawford and Kuperman 2006; Kuperman 2008a, 2008b).\(^1\)

Empirically, the debate focuses on the post–Cold War period during which support for intervention has grown and humanitarian intervention has become more frequent. Bellamy and Williams (2012) point out that the incidence and duration of civil wars and mass atrocities have declined over the past 20 years, as the norm supporting intervention has taken hold, which seems in conflict with the moral hazard thesis.\(^2\) Hultman (2010) finds that...
UN interventions mandated to protect civilians do reduce civilian deaths, but other UN interventions increase rebel targeting of civilians. In support of the moral hazard perspective, scholars such as Kuperman (2008b) and Grigoryan (2010) present detailed case studies of recent conflicts, especially Bosnia, Kosovo, and Darfur. Kosovo Liberation Army (KLA) leaders openly acknowledged that they would lose without intervention but hoped to provoke Serbian atrocities in order to draw in the international community.

At the theoretical level, the arguments presented by each side are incomplete and fail to fully grasp the strategic dynamics of intervention and civil wars. Advocates of intervention adopt a deterrence perspective and focus on two actors, the third-party intervenor and the potential perpetrator. Credible threats by the third party are held to deter the perpetrator, solving the problem. But this ignores the provocative effect of intervention on the rebel group, as moral hazard critics claim. But the moral hazard position in turn ignores the state side and has a hard time explaining why, if intervention makes the rebel group more demanding, it does not make the state more conciliatory, thereby canceling out the effect. Any adequate model of the situation must consider three fully strategic actors: the state, the rebel group, and the third-party intervenor.

We present a model of intervention in which a government and rebel group bargain and may fight a civil war if they fail to reach agreement. If they fight, the government decides what level of atrocities to commit, and a third party may intervene to halt atrocities. By including a number of different parameters, the model underscores how complex the intervention conundrum actually is. Concretely, the model yields the following findings. First, humanitarian intervention will typically strengthen the rebel group and weaken the state. If the parties are risk neutral, these effects cancel out, and war is no more likely than without intervention. However, if the parties are risk acceptant, and if intervention creates a balance of power between the two, then the rebel group may become more demanding than the state becomes conciliatory. In this case, a more neutral intervention that does not side as much with the rebel group is more likely to minimize atrocities. Second, intervention to limit atrocities, if successful in enforcing that limit, reduces the cost of war, which makes it more likely. These effects may also cancel out, which would leave the expected level of atrocities unchanged. However, it would be better to make war less costly and no more likely. The model shows that if compensating costs are imposed on the leaders to make up for the fact that war is less costly for their citizens, then the cost of war to citizens can be reduced without increasing the likelihood of war.

On balance, therefore, we find that a well-crafted intervention regime can reduce the expected level of atrocities and so be counted as welfare improving overall. The dangers pointed out by the moral hazard critique are real but can be compensated for by proper consideration of the logic of intervention revealed by the model. In what follows, we first discuss logic of the problem in more detail, then present our model, then discuss its implications and conclude.

The Logic of Humanitarian Intervention

The question we address is whether humanitarian intervention will reduce the expected level of atrocities committed by states in civil wars. By atrocities we mean large-scale, sustained, intentional, systematic violence against civilians that results in mass deaths through murder or the deliberate creation of conditions in which civilians die. Some prominent recent cases are Bosnia, Rwanda, Kosovo, East Timor, Darfur, the Democratic Republic of Congo, Libya, and Syria. The outcome we have in mind is consistent with the term “mass atrocities,” a term that includes genocide, ethnic cleansing, crimes against humanity, and war crimes and that has become the term of reference in the policy community (Evans 2008; Sewell, Raymond, and Chin 2010). Mass atrocities are most common in the context of armed conflict, particularly civil war (Straus 2007; Ulfelder and Valentino 2008). While we recognize that nonstate actors or rebel groups engage in atrocities, we focus on state-perpetrated or state-led atrocities. Such atrocities are the most common—states generally have the resources to commit large-scale, sustained violence against civilians—and have been the main focus of international attention. By expected level of atrocities we mean the product of the level of atrocities per war and likelihood of war. What is important is not how many people die per war, or how many wars there are per year, but how many people die from war per year.

We assume that actors weigh the costs and benefits when deciding to fight and when they commit atrocities (Kalyvas 2006; Valentino 2004; Valentino, Huth, and Balch-Lindsay 2004). Intervention changes the actors’ calculations by altering their payoff for war. The war payoff is a function of (1) the probability distribution over the possible outcomes of the war and (2) the costs of conflict. Intervention can alter the distribution over outcomes by...
helping one side to achieve its objectives, making it more likely to win and its opponent less likely. Intervention can also affect the costs of war by preventing atrocities that would have otherwise occurred and punishing actors that commit them. These effects can in turn influence whether the parties fight and if so with what level of atrocities.

Advocates of intervention argue that credible threats of third-party intervention to stop atrocities should have the intended effect by making it more costly and less useful for states to commit atrocities. If they intervene, third parties will impose additional costs on the state for committing atrocities, and shift the distribution over outcomes against it, in favor of the rebel group. Therefore, in the presence of an intervention regime, states should be more reluctant to commit atrocities. The main problem, in this view, is in making the threats credible. For instance, Harvey analyzes the Bosnia conflict and finds that early threats lacked credibility (Harvey 1997). Scholars also explore how humanitarian intervention can be squared with international law and the rights of sovereignty and what are the mechanics of doing it properly (Chalk et al. 2010; Cooper and Kohler 2009; Evans 2008; Hamburg 2010; Sewell, Raymond, and Chin 2010).

Critics of intervention argue that intervention may actually make atrocities more likely by encouraging rebellion. The prospect of intervention may encourage rebellion by strengthening the hand of rebel groups (shifting the distribution over outcomes in their favor) and making war less costly, hence more attractive (Crawford and Kuperman 2006; Jenne 2004; Rowlands and Carment 1998). Kuperman argues this dynamic was at work in the Kosovo conflict in 1999 in that the Kosovo rebels realized they were too weak to successfully confront Serbia by themselves but hoped to provoke massacres which would bring in the international community (Kuperman 2008b). The logic of moral hazard holds that the prospect of international intervention on behalf of weak victim groups encourages risky rebellions, much as insurance is said to encourage greater risk taking. The usual remedies for this kind of problem in economic contexts are difficult to apply in the intervention context, but Kuperman (2008a) recommends close attention to whether rebel groups provoke state genocides or are genuinely responding to them, and only intervening in the latter case.

Each of these perspectives is an important, but partial, take on the subject. In particular, two trade-offs require more careful investigation. First, if intervention strengthens the rebel group at the expense of the state, the rebels should become more demanding and the state should become more conciliatory (Grigoryan 2006; Rauchhaus 2006). What will be the net effect on the prospect of war? Under what conditions will the rebels demand more than the state is willing to concede, and can the third party make this less likely? Second, if intervention encourages war by lowering its cost, does this produce more frequent but less costly wars? Relatedly, what if the prospect of intervention causes some states to escalate their level of atrocities in war (Grigoryan 2010)? Will this make war less likely? Can anything be done to make wars less costly and less frequent?

The game-theoretic literature has not yet provided convincing answers to these questions. Cetinyan (2002) argues that the prospect of third-party intervention should have no impact on the likelihood of conflict, only on the terms of the agreement. However, he models war as a deterministic result of the balance of power and effort between the third party and the state actor, when it is usually considered a risky gamble. The closest analysis to ours is by Grigoryan (2010), who examines escalation by the state to higher levels of violence within a civil conflict. Escalation occurs when there is uncertainty about whether the third party will intervene. While very useful, Grigoryan makes two assumptions that are not quite appropriate for our purposes. First, he assumes that the demand made by the minority group is fixed, rather than adjustable. Given that issue indivisibilities can cause conflict as well as uncertainty, it is unclear which is driving the conflict. We assume instead that the parties can make any demand along a continuum and so can smoothly adjust the demand in response to changes in the model parameters. Second, the way Grigoryan structures the model, escalation to high levels of violence can only happen after an initial intervention by the third party. While this closely corresponds to the Kosovo case, it ignores the fact that many states commit atrocities without any preliminary intervention. A model that fairly evaluates the utility of intervention must account for such cases as well.

A Model of Intervention

There are three actors: player 1, the government side; player 2, the rebel side; and the third party and potential

---


4Rauchhaus (2009) points out that third parties’ ability to monitor rebel-group behavior means that there is no “hidden action,” which is a foundation of moral hazard models.

5Two other related models are Yuen (2009) and Benson et al. (2012).
FIGURE 1 The Intervention Game

\[
\begin{tikzpicture}
    \node (player1) at (0,0) {
        Player 1
    };
    \node (player2) at (-2,-2) {
        Player 2
    };
    \node (player3) at (2,-2) {
        Player 3
    };
    \node (offer) at (0,-4) {
        Offer \(x\)
    };
    \node (accept) at (-2,-4) {
        Accept
    };
    \node (reject) at (2,-4) {
        Reject
    };
    \node (choose) at (0,-6) {
        Choose atrocity level \(a\)
    };
    \node (impose) at (0,-8) {
        Impose Sanctions
    };
    \node (intervene) at (0,-10) {
        Intervene
    };

    \draw[->] (player1) -- (offer);
    \draw[->] (player1) -- (choose);
    \draw[->] (player1) -- (impose);
    \draw[->] (player1) -- (intervene);
    \draw[->] (player2) -- (offer);
    \draw[->] (player2) -- (choose);
    \draw[->] (player2) -- (impose);
    \draw[->] (player2) -- (intervene);
    \draw[->] (player3) -- (offer);
    \draw[->] (player3) -- (choose);
    \draw[->] (player3) -- (impose);
    \draw[->] (player3) -- (intervene);

\end{tikzpicture}
\]

intervenor, player 3. The players bargain over an issue represented by the unit interval \(X = [0, 1]\). This could represent the level of central authority enjoyed by the state, where \(x = 0\) represents complete independence for player 2 and \(x = 1\) represents a centralized state controlled by player 1. Players 1 and 2 have strictly opposed preferences in that player 1 wants \(x\) to be higher, and player 2 wants it to be lower. Player 3 most prefers some outcome in the interval, \(I_3 \in X\), and likes outcomes less the further they are from its ideal point. Players 1 and 2 value their own ideal point at 1 and their adversary’s at 0, \(u_i(I_1) = 1\), \(u_i(I_2) = 0\). The third party also values its own ideal point at 1 and the other players’ not less than zero, \(u_i(I_3) = 1\), \(u_3(0) = u_3(1) \geq 0\). In the case of repressive authoritarian regimes such as Libya under Qadaffi, the third party will often favor the rebel side. State atrocities will only further tilt the third party in favor of the rebel group. In these cases, \(I_3\) is closer to 0 than 1, and \(u_3(0) \geq u_3(1)\).

The three players play a game illustrated in Figure 1. Player 1 starts the game by making an offer to player 2 of some issue resolution \(x \in X\). If player 2 accepts the offer, the game ends with payoffs \(u_1(x), u_2(x), u_3(x)\). If player 2 rejects the offer, a war breaks out. Player 1 then has the choice of how to fight the war. Specifically, it chooses a level of atrocities to commit, \(a \in [0, 1]\) where 0 indicates a “clean” counterinsurgency campaign that scrupulously avoids targeting noncombatants, and 1 indicates an all-out genocidal campaign against civilians associated with the rebels.\(^7\) Player 3 then has a choice to impose sanctions on the party committing atrocities or to intervene militarily in the war. We assume that sanctions remain in place even if there is an intervention, so the choice is to impose sanctions alone or in combination with intervention. Sanctions impose a cost \(s \geq 0\) on player 1.

The payoffs for war are shown in Table 1. We model war, following the usual convention, as a lottery in which the victor imposes its ideal point (Powell 2002). Departing slightly from convention, we assume that this holds true in a trilateral war as well as a bilateral one. That is, we assume that the third party may “win,” and if it does, it imposes \(I_3\) as the outcome. If the third party’s ideal point is 1 or 0, it will effectively fight on behalf of the side it prefers. If its ideal point is more neutral, for instance \(I_3 \approx 0.5\), it will prefer, and fight to achieve, a compromise solution. The model therefore covers both biased and neutral third parties.

We assume that the government’s chance of victory against the rebels may depend on the level of atrocities. Specifically, we assume that player 1’s chance of winning a bilateral war is

\[
p_1^b = \alpha + \beta a
\]

where \(\alpha, \alpha + \beta \in (0, 1)\). The marginal benefit of atrocities is \(\beta\). If \(\beta > 0\), committing more atrocities increases the chance of victory against an insurgent group (Valentino, Huth, and Balch-Lindsay 2004). However, atrocities may not help or hurt a perpetrator’s war effort, in which case \(\beta = 0\). It may even be the case that committing more atrocities decreases one’s chance of victory, in which case \(\beta < 0\). We define the rebels’ chance of winning a bilateral war as \(p_2^b = 1 - p_1^b\). We assume that the third party’s chance of winning if it decides to intervene, denoted \(p_3\), is unaffected by the level of atrocities.\(^8\) The

\(^7\) We acknowledge that in reality the decision on what level of atrocities to commit is often made incrementally as the war progresses. Our model can be interpreted as covering the portion of this process where the state chooses to remain at a level of atrocities that would not draw international intervention or escalate to a level that would.

\(^8\) We make this assumption because the third party does not depend on the victimized population for cover and support, as the rebel group does. The third party’s military forces are usually well armed.
chance of winning a trilateral war after an intervention for the government and rebels is defined as $p_i^*(1 - p_3) p_i^*$.

The parties pay costs if there is a conflict. The direct cost for player $i$ of fighting is denoted $c_i$. The third party only incurs direct costs if it decides to intervene. The level of atrocities also affects the players; $k_ia$ is subtracted from each player’s payoff. In the case of the rebel group and the third party, we assume that $k_1 > 0$, so greater atrocities are felt as a greater cost. In the case of the government, we allow $k_1$ to be positive or negative. A positive $k_1$ corresponds to a government that finds it intrinsically costly to commit atrocities and would therefore only commit them to improve its chance of victory. We say that such a government is instrumental in its attitude toward atrocities. A negative $k_1$ corresponds to a government that derives intrinsic utility from atrocities aside from any strategic value they may have. Such a player could engage in atrocities even if they harm the war effort, if the intrinsic benefit is great enough. The costs from atrocities are incurred by the third party regardless of whether it intervenes. Finally, if the third party does intervene, it punishes player 1 for committing atrocities by imposing an additional cost $\kappa$. This represents the additional sanctions and other punishments that accompany interventions targeted at coercing a state into refraining from atrocities, such as the punishment of leaders involved in atrocities before international tribunals.

Finally, we assume that when they intervene, third parties prevent some atrocities that would otherwise have taken place.9 That is, we assume that intervention reduces the level of atrocities by a factor of $\delta \in (0, 1)$ so that if the government selects a level $a$ and the third party intervenes, the resulting level of atrocities will be $\delta a$. This is not to deny that the government might escalate to high levels of atrocities in anticipation of an intervention designed to prevent atrocities; indeed, this is a central topic below.

### Equilibria with Complete Information

We first solve the game assuming complete information between the two bargaining parties, using subgame perfection as the solution concept. That is, the actors’ payoffs are common knowledge among the players, and all actions must be rational at the time they are carried out.

Starting at the end of the game, the third party will intervene if $p_i^*(a) u_i(1) + p_i^*(a) u_i(0) - k_3a < p_i^*(\delta a) u_i(1) + p_i^*(\delta a) u_i(0) + p_3 - c_3 - k_3\delta a$, which implies that it will intervene if the level of atrocities rises above the following threshold:10

$$a^* = \frac{c_3 - p_3(1 - \alpha u_i(1) - (1 - \alpha) u_i(0))}{(1 - \delta)k_3 + \beta(1 - (1 - p_3)\delta)(u_i(0) - u_i(1))}$$ (1)

Player 3 wants to stop atrocities and will intervene if their level rises above the threshold $a^*$. By intervening it can prevent some atrocities, as well as make its preferred war outcome more likely.

How do the parameters affect this threshold? Increasing the cost of intervention, $c_3$, raises the threshold, so player 1 can commit more atrocities without triggering intervention. Increasing player 3’s sensitivity to atrocities, $k_3$, will decrease the threshold, constraining player 1.

### Table 1 The War Payoffs

<table>
<thead>
<tr>
<th></th>
<th>Without Intervention</th>
<th>With Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Player 1</td>
<td>$p_i^*(a) - c_1 - k_1a - s$</td>
<td>$p_i^*(\delta a) + p_3u_i(1) - c_1 - k_1\delta a - s - \kappa$</td>
</tr>
<tr>
<td>Player 2</td>
<td>$p_i^*(a) - c_2 - k_2a$</td>
<td>$p_i^*(\delta a) + p_3u_i(1) - c_2 - k_2\delta a$</td>
</tr>
<tr>
<td>Player 3</td>
<td>$p_i^<em>(a)u_i(1) + p_i^</em>(a)u_i(0) - k_3a$</td>
<td>$p_i^<em>(\delta a)u_i(1) + p_i^</em>(\delta a)u_i(0) + p_3 - c_3 - k_3\delta a$</td>
</tr>
</tbody>
</table>

---

9 We base this assumption on an apparent consensus in the humanitarian intervention literature that well-crafted interventions with clear mandates to protect civilians—in other words, the type of action that the intervention regime imagined here would institutionalize—save lives. Based on an analysis of 17 cases in the 1990s, Seybolt (2007) finds that well-designed, rapidly deployed interventions with civilian protection mandates were effective. Arguing that humanitarian interventions were infrequent and inconsistently chosen in the 2000s, Weiss (2012) points to actions in Libya and Côte d’Ivoire as examples where international interventions saved lives, a finding also echoed in Bellamy and Williams (2011). At the same time, we recognize that empirical studies measuring the effectiveness of interventions remain underdeveloped,

10 In equilibrium, player 3 must break ties in favor of not intervening.
If player 3 is biased in favor of the rebel group, so that $u_3(0) > u_3(1)$, and atrocities help player 1’s war effort, $\beta > 0$, then increasing the power of the third party, $p_3$, and decreasing the level of postintervention atrocities, $\delta$, decreases the threshold, reducing the level of atrocities that will be possible without intervention. Also in this case, increasing the effect of atrocities on the likelihood of winning, $\beta$, will lower the threshold, since it gives player 3 extra reason to intervene on behalf of the side it favors. Increasing $\alpha$, which effectively makes player 1 stronger, also lowers the intervention threshold.\footnote{Note, player 3 may intervene even if it values the two side’s ideal points the same, $u_3(0) = u_3(1)$, and even if it values them the same as its own ideal point, $u_3(0) = u_3(1) = 1$, which implies that it does not care at all about the outcome. Thus very neutral third parties may still intervene.}

This threshold may be less than zero or greater than 1. That is, the third party might be willing to intervene even in the absence of atrocities or unwilling to intervene even with maximal atrocities. If the following condition holds, the level of atrocities chosen will affect the prospect of intervention:

$$a^1 \in [0, 1).$$

(2)

In the text, we focus on this case where equation 2 holds, and the supporting information discusses the cases in which the level of atrocities does not affect the prospect of intervention.

Turning to player 1’s choice of the level of atrocities in the bilateral case, it is found by taking the derivative of its war payoff with respect to the level of atrocities. In the bilateral case, this is $2p^1(a) - k_1$. Therefore, player 1 will find it worthwhile to commit atrocities in the bilateral context if the marginal benefit exceeds the marginal cost,

$$\beta > k_1.$$ (3)

If intervention is expected, the marginal benefit is reduced because the third party’s chance of winning is not affected by committing atrocities, so the condition is the following:

$$(1 - p_3)\beta > k_1.$$ (4)

Note these conditions could hold even if atrocities are costly to player 1 ($k_1 > 0$), so long as their positive impact on the prospect of winning exceeds their cost. If atrocities hurt player 1’s prospects of winning ($\beta < 0$), these conditions can still be satisfied so long as they are sufficiently intrinsically valued ($k_1 < 0$). In the case where intervention is expected, the more powerful the intervention (high $p_3$) the less impact atrocities have on the distribution over outcomes, so if they are costly ($k_1 > 0$), they will be avoided, and if they are intrinsically valued ($k_1 < 0$), they will be committed. In the text, we consider the case where equations (3) and (4) are satisfied and so atrocities are worth committing, and we discuss the alternative cases in the supporting information.

Assuming that equations (2), (3), and (4) are satisfied so the level of atrocities affects the prospects of intervention and atrocities are worth committing, any level $a < a^1$ is dominated by $a^1$ since the latter results in no intervention and a higher utility. Any level greater than $a^1$ but less than 1 is dominated by 1 because intervention is assured, so player 1 might as well maximize atrocities. Player 1 therefore compares the utility of choosing $a^1$, which results in no intervention, to choosing $a = 1$, which provokes intervention. Player 1 will choose to abide by the threshold if

$$p^1(a^1) - c_1 - k_1a^1 - s \geq p^1(\delta) + p_3u_1(I_s) - c_1 - k_1\delta - s - \kappa,$$

or,

$$p^1(a^1) - p^1(\delta) + p_3(p^1(\delta) - u_1(I_s)) + k_1(\delta - a^1) + \kappa \geq 0.$$ (5)

If equation (5) holds, then player 1 will abide by the threshold so the equilibrium level of atrocities will be the threshold level, $a^* = a^1$. If not, then player 1 will escalate to all-out atrocities, and the equilibrium level will be $a^* = 1$. The greater the threshold, $a^1$, the more likely equation (5) will hold and player 1 will observe the threshold.

Turning now to player 2’s choice, there are two cases to consider. If player 1 will choose $a^1$ (equation (5) is satisfied), the war will be bilateral and player 2 will accept the offer if it is better than war,

$$u_2(x) \geq p^1(a^1) - c_2 - k_2a^1,$$

and reject it otherwise. Let $r^2_x$ denote the highest value of $x \in [0, 1]$ that satisfies this condition, that is, the worst deal that player 2 will accept rather than fight, or its reservation value. If equation (5) is not satisfied, then player 1 opts for $a = 1$, there will be intervention, and player 2 will accept the offer if it beats the multilateral war payoff,

$$u_2(x) \geq p^1(\delta) + p_3u_2(I_3) - c_2 - k_2\delta,$$

and reject it otherwise. Let $r^2_x$ be the highest value of $x \in [0, 1]$ that satisfies this condition.

Finally, consider player 1’s offer. If the war will be bilateral, player 1 compares player 2’s bilateral reservation value with its own war payoff and will make the minimal offer that satisfies player 2 if it is better for player 1 than war, or if

$$u_1(r^2_x) \geq p^1(a^1) - c_1 - k_1a^1 - s.$$
If equation (5) does not hold and intervention will take place, player 1 will make the peaceful offer if

\[ u_1(r_1^*) \geq p_1(\bar{\delta}) + p_3 u_1(I_3) - c_1 - k_1 \bar{\delta} - s - \kappa. \]

If we denote player 1’s reservation values in the bilateral and trilateral cases, or the lowest value of \( x \in [0, 1] \) that makes it at least well off as war, as \( r_1^b \) and \( r_1^t \) respectively, then peace will be possible in the bilateral case if \( r_1^b \leq r_1^t \) and in the trilateral case if \( r_1^t \leq r_1^t \).

Summing up, in the subgame perfect equilibrium, player 3 will intervene if player 1 goes beyond the threshold, \( a > a^1 \). Player 1 will abide by the threshold if equation (5) holds and will escalate to \( a = 1 \) otherwise. Player 2 sets its minimum acceptable offer, or reservation value, based on the nature of the anticipated war. Player 1 will offer player 2 its reservation value if it prefers that outcome to war and otherwise will make an unacceptable offer, precipitating a conflict.

We show in the supporting information that with risk-neutral preferences, there will be no war in the complete information version of the game, regardless of whether or not intervention takes place. However, with risk-acceptant preferences there can be war, and so intervention can make war more or less likely. We discuss this case in the implications section below.

**Equilibria with Incomplete Information**

We introduce uncertainty over player 2’s costs for fighting. Assume that player 2’s direct costs for fighting are uniformly distributed over the interval \([0, 1]\). This assumption ensures that the highest cost type will accept any deal at all rather than fight while the lowest cost type will fight so long as the costs of atrocities are not too great. The likelihood that player 2 has costs below \( c_2 \) is \( F_2(c_2) = c_2 \). Player 2 learns its own cost when the game begins, and the other players are uncertain.

The intervention decision by the third party is the same as in the complete information case, as is the decision by player 1 on the level of atrocities to select. When we turn to player 2’s decision, however, there are now a continuum of types and a resulting probability of acceptance and rejection. We focus again on the case where equations (2), (3), and (4) are satisfied so the level of atrocities can affect the prospect of intervention and atrocities are worth committing.

If player 1 will choose \( a^1 \), the war will be bilateral, and player 2 will reject the offer if \( c_2 < p_2^b(a^1) - k_2 a^1 - u_2(x) \). The probability of bilateral war, denoted \( P(W^b) \), is just the likelihood that the offer is rejected, which, given the uniform distribution of \( c_2 \) on \([0, 1]\), is the following:\(^{12}\)

\[ P(W^b) = \max \left\{ p_2^b(a^1) - k_2 a^1 - u_2(x), 0 \right\} \]

If player 1 opts for \( a = 1 \), there will be intervention, and player 2 will reject the offer if \( c_2 < p_2(\bar{\delta}) + p_3 u_2(I_3) - k_2 \bar{\delta} - u_2(x) \), and the likelihood of trilateral war in this case is

\[ P(W^t) = \max \left\{ p_2^b(\bar{\delta}) + p_3 u_2(I_3) - k_2 \bar{\delta} - u_2(x), 0 \right\}. \]

**The Equilibrium Offer**

Once again we consider two cases, one in which player 1 will opt for \( a^1 \) and one in which it will choose \( a = 1 \). When player 1 will choose \( a^1 \), the payoff for offering \( x \) is the probability of war times the war payoff, plus the probability of peace times the value of the offer:

\[ P(W^t)(p_1^b(a^1) - c_1 - k_1 a^1 - s) + (1 - P(W^b)) u_1(x). \]

When player 1 will choose \( a = 1 \), the war will be trilateral and the payoff for offering \( x \) is

\[ P(W^t)(p_1(\bar{\delta}) + p_3 u_2(I_3) - c_1 - k_1 \bar{\delta} - s - \kappa) + (1 - P(W^t)) u_1(x). \]

Player 1 will choose an offer, denoted \( x^* \), that maximizes its payoff, depending on its future behavior in war. Explicit solutions can be easily derived if we assume risk-neutral preferences. We focus on interior solutions, in which the demand is less than 1, and there is some chance of war. Other cases and algebra are in the supporting information.

If player 1 will observe the threshold and set \( a = a^1 \), the equilibrium offer that player 1 chooses is

\[ x^* = p_1^b(a^1) + \frac{1 + (k_2 - k_1)a^1 - c_1 - s}{2}. \]

If player 1 will not observe the threshold and set \( a = 1 \), the equilibrium offer in the case of intervention is as follows:

\[ x^* = (1 - p_1) p_1^b(\bar{\delta}) + p_3 I_3 + \frac{1 + (k_2 - k_1)\bar{\delta} - c_1 - s - \kappa}{2}. \]

We discuss comparative statics on the equilibrium demand in the supporting information.

\(^{12}\)The expression could be negative, hence the need to truncate it at zero. It cannot be greater than 1.
The Probability of War

If player 1 will obey the threshold and the war will remain bilateral, then the equilibrium likelihood of war is

$$P(W^b) = \frac{1 - (k_1 + k_2) a^\dagger - c_1 - s}{2}. \quad (6)$$

If player 1 will escalate to $a = 1$ and the third party intervenes, the probability of war is

$$P(W^t) = \frac{1 - (k_1 + k_2) \delta - c_1 - s - \kappa}{2}. \quad (7)$$

Note that these are identical except for two factors. First, player 1’s additional punishment from intervention, $\kappa$, enters negatively in the trilateral case, lowering it in comparison with the bilateral case. Second, $\delta$ replaces $a^\dagger$. The comparative statics are straightforward. Anything that increases the cost of war makes war less likely. Increasing player 1’s costs for combat, sanctions, and intervention, $c_1$, $s$, and $\kappa$, decreases probability of war. Increasing the players’ costs for atrocities, $k_1$ and $k_2$, also decreases the likelihood of war.

The impact of the threshold of intervention, $a^\dagger$, and the postintervention level of atrocities, $\delta$, depends on the sign of $k_1 + k_2$. If this is positive,

$$k_1 + k_2 \geq 0, \quad (8)$$

then increasing the level of atrocities lowers the probability of war. We consider this to be the usual case; if atrocities are felt as a joint cost, then increasing them lowers the likelihood of war and decreasing them raises the likelihood of war. Note that if player 1 is instrumental in its attitude towards atrocities, then $k_1 > 0$ and equation (8) is satisfied. However, if equation (8) is not satisfied, because, bluntly, player 1 really wants to kill player 2’s civilians, and player 2 does not care much if they are killed, then increasing the level of atrocities will increase the likelihood of war because player 1’s war payoff increases more than player 2’s decreases.

The Equilibrium Level of Atrocities

The expected level of atrocities is the probability of war times the level of atrocities. In the bilateral case, this is

$$E(a | W^b) = \frac{1 - (k_1 + k_2) a^\dagger - c_1 - s}{2} a^\dagger,$$

and in the trilateral case, it is

$$E(a | W^t) = \frac{1 - (k_1 + k_2) \delta - c_1 - s - \kappa}{2} \delta.$$  

The impact of the costs operates through the probability of war and is therefore the same as previously. If the players’ costs for combat, atrocities, sanctions, and intervention, $c_1$, $k_1$, $k_2$, $s$, and $\kappa$, increase, then the expected level of atrocities goes down. These variables reduce the likelihood of war without having any countervailing impact on the level of atrocities per war, so their impact is unambiguous.

Changing the threshold level of atrocities, $a^\dagger$, and the level of postintervention atrocities, $\delta$, has a more complicated effect. Lowering the level of atrocities makes each war that occurs less costly. However, if equation (8) is satisfied, lowering the level of atrocities makes war more likely. These effects work at cross purposes.

Implications and Analysis

We focus on the two questions about the impact of intervention discussed earlier. First, does strengthening the rebel side make war more likely, and if so under what conditions? Second, does reducing the cost of war for civilians make war more likely, and if so are there alternative policies that can reduce the civilian cost of war without raising its likelihood?

Strengthening the Rebel Side

One of the main questions left unanswered by the existing literature is whether intervention, by strengthening the rebel side and weakening the state, makes war more likely by encouraging the rebels to be more demanding, or whether the corresponding weakening of the state position sufficiently counteracts this. Advocates of intervention focus on how intervention weakens the state, and proponents of moral hazard focus on how intervention strengthens the rebel group.

We show in the supporting information that if the parties are risk neutral, these effects do indeed cancel out. If the parties are risk neutral, altering the balance of power between them, or more generally the distribution over outcomes, will have no effect on the incidence of war. An intervention regime that strengthens rebel groups and weakens the state will just shift the bargaining range in favor of the rebels, without provoking war. Cetinyan’s result is therefore supported in the risk-neutral case.

However, with risk-acceptant preferences, these effects may not cancel out. Risk-acceptant preferences devalue compromise solutions in favor of all-or-nothing gambles. Faced with a choice between achieving 50% of what is sought and going to war with a 50% chance of winning, the risk-acceptant actor chooses war, if the costs
are not too great. With risk-acceptant preferences, imbalances of power are conducive to peace while balances of power can lead to conflict. When power is very imbalanced, the weaker actor, even if risk acceptant, will be willing to live with very little, given the low chance of victory. When power is balanced, however, each side has a nonnegligible chance of winning, yet the compromise solutions that would otherwise be acceptable to both sides are undervalued. The danger, therefore, is that in situations where the state is strong, or would be if unconstrained, intervention may strengthen the rebel group and produce a balance of power where there previously was imbalance and that this may lead to war.

Intervention may produce a balance of power in two ways. First, it could be that intervention produces a balance of power, but player 1 ignores the threshold and triggers intervention anyway because of the utility of atrocities and the inability of the intervention to suppress them altogether. Assume the third party is biased in favor of the rebel group (low $I_3$) but not all powerful (low $p_3$); this will mean that the rebel's ideal point becomes more likely to be realized if the third party intervenes, but not certain. If atrocities are helpful to the war effort (high $\beta$), and the third party is unable to greatly suppress them when they intervene (high $\delta$), then player 1 will ignore the threshold and trigger intervention. However, since the resulting war will exhibit a balance of power, the players will not be willing to compromise in advance, so war will follow. In this case, the actual intervention produces the balance of power and causes war.

Alternatively, even the prospect of intervention, though not realized in equilibrium, may cause war. Consider a case where the power balance between players 1 and 2 would be roughly balanced if there were no atrocities but very imbalanced with high atrocities ($\alpha \approx \beta \approx 0.5$). Left to their own devices, player 1 would threaten genocidal war, which would convince player 2 to accept very little in the bargaining. Now let there be a very effective intervention regime, powerful (high $p_3$) and capable (low $\delta$), and therefore able to impose a low threshold on atrocities (low $a'\delta$). Player 1 therefore decides to obey the low threshold in equilibrium to avoid intervention. However, with low levels of atrocities, there is a balance of power, so the two sides' demands are now irreconcilable, and bilateral war results. The third party never intervenes, but the intervention regime has nonetheless caused the war.

One factor that may affect the tendency of intervention to produce power balances is the third party's ideal point. In particular, if the third party is relatively neutral ($I_3 \approx 0.5$), a victory for the third party will produce an outcome which corresponds to neither side's ideal point. In particular, player 2 will derive little satisfaction from the intervention of such a third party, so they will not become as intransigent in the bargaining if intervention is expected. By reducing the level of atrocities, intervention may still benefit player 2 in the bargaining to some degree, but not sufficiently to encourage it to provoke a war. The result will be peace on slightly better terms for player 2 than in the absence of intervention. Neutral third parties may therefore be somewhat less likely to provoke war than ones biased in favor of the rebel side.

The moral hazard perspective is correct in arguing that intervention can provoke war by strengthening the weaker side, but this occurs under specific conditions. When power is imbalanced, and the third party is biased in favor of the weaker side, it may help move the balance of power into the middling range, which makes compromise impossible if the players have risk-acceptant preferences. However, this problem is less severe when the third party has a more neutral ideal point and does not favor one side or the other. A general lesson is that intervenors should avoid creating balances of power, either by limiting atrocities to a level at which power between the two sides is balanced in a bilateral war or by intervening half-heartedly on behalf of a weaker side so that the resulting trilateral conflict is essentially balanced between two opposing camps.

### Making War Less Costly

The second main question left unanswered by the literature on intervention is whether reducing the level of atrocities makes war more likely, what the net effect is on the expected level of atrocities, and what can be done to minimize this quantity. We consider the cost-related impact of intervention in the incomplete information version of the model by examining the effect of lowering the threshold for atrocities on the expected level of atrocities. If $a'$ is lowered (because of a change in $k_3$ or $c_3$ which does not affect any other calculations), there will be three types of state to consider: those who were not obeying the old threshold and will not obey the new one, those who were obeying the old threshold and decide to obey the new one, and those who switch from obeying the threshold to violating it.

Those who were not obeying the original threshold will still fight trilateral wars. Nothing will change for this population, so there will be no impact on the expected level of atrocities in this case. Those who continue to obey the threshold are still fighting bilateral wars. Their level of atrocities goes down, and the likelihood of war goes up, if player 1 is instrumental so equation (8) is satisfied. For these cases, lowering atrocities raises the likelihood of war.
Finally, some will switch from obeying the threshold to not obeying it, in accordance with equation (5). For these types, war that was once bilateral is now trilateral. The level of atrocities per war goes up from the old threshold to the postintervention level of atrocities, $\delta$. The likelihood of war goes down, assuming instrumental players, since player 1 suffers the additional cost of intervention, $\kappa$, and player 2 suffers the cost of the increased level of atrocities. The effect on the expected level of atrocities is therefore mixed, and in the opposite way from the previous comparison. Lowering the threshold raises the level of atrocities, which lowers the probability of war.

What, then, is the net effect of lowering the intervention threshold, $a^1$, on the expected level of atrocities? Unfortunately, there is no straightforward answer; it depends on the specific values of a variety of parameters and the empirical frequency of the different types of states. Lowering the threshold of intervention may raise or lower the ultimate expected level of atrocities.

This indeterminacy suggests the need to consider a broader set of policy tools that could potentially be manipulated together in hopes of lowering the level of atrocities per war without making war more likely. First, consider the fact that lowering the threshold of intervention can cause some actors that previously abided by the threshold to ignore the new threshold and escalate to $a = 1$. Can this be prevented by manipulating other variables? If we examine equation (5), we can see that there are three variables that can be manipulated to keep the expression constant, despite lowering $a^1$. First, if the postintervention level of atrocities, $\delta$, is lowered, this will raise the left-hand side of equation (5). Second, since player 1 dislikes intervention ($p^1(\delta) > u_I(I_1)$), increasing the strength of the intervention, $p_3$, will also raise the expression. Finally, increasing the cost imposed on player 1 for committing atrocities, $\kappa$, will also raise the expression, making player 1 more likely to observe the threshold.

Note that $\delta$ and $p_3$ enter into both the expression for $a^1$ (equation (1)) and equation (5). Making the intervention stronger by increasing $p_3$, and better able to prevent atrocities by lowering $\delta$, will both lower the threshold of intervention and increase the expression in equation (5), counterbalancing the effect of lowering $a^1$ and making player 1 more likely to obey the new threshold. However, these effects might not balance out exactly, which could lead a marginal player 1 to change its decision. To deal with this problem, we note that player 1’s cost for intervention, $\kappa$, appears in equation (5) but does not appear in equation (1), so it can be used to adjust equation (5) at will to compensate for changes in $a^1$. That is, one can in theory find an increased level of punishment for states that commit atrocities sufficient to convince them if they were willing to abide by the threshold limit on atrocities before, to continue to do so even if that threshold is lowered. This is particularly true if the threshold reduction has been accomplished by making the potential intervention stronger and more capable of reducing atrocities.

Now consider the probability of war. The crux of the objection to lowering the threshold for intervention is that it raises the probability of war. Can an alternative variable be found to negate this side effect? The sanctions variable, $s$, is the obvious candidate. By increasing the level of sanctions, the international community can make war more costly for player 1. If we examine equations (6) and (7), it is easy to see that the positive effect on the probability of war of reducing $a^1$ can be compensated for by increasing $s$ so that the probability of war remains the same. In fact, one could argue that this kind of change is itself normatively desirable, particularly if the sanctions are “smart” and target the leaders with travel restrictions, financial barriers, and the like. Transferring the costs of civil war from the civilian victims of atrocities to the bank accounts of the leaders is all to the good, and all the more practical since leaders may often care more about their own well-being than that of their followers.

The bottom line is that a strengthened intervention regime can both deter states from escalating past the newly lowered threshold and convince states to make acceptable offers by being able to intervene forcefully and effectively and by imposing alternative costs that keep war costly for leaders while it is made less costly for their followers.

**Conclusion**

Large-scale violence against civilians is a problem that unfortunately remains pressing, as the recent cases of Darfur, Libya, Côte d’Ivoire, and Syria make clear. International actors face pressure to develop policies and institutions to counter mass atrocities. But scholars remain divided about whether an antiatrocities regime will lower the amount of atrocities worldwide or have the opposite effect. Existing arguments are partial and simplistic, either clamoring for the political will to deter perpetrators and save victims or condemning naive humanitarianism that provokes what it seeks to prevent.

While any model simplifies reality, our model seeks to deepen our understanding of the intervention conundrum by including all three relevant actors—the state, rebels, and third party—and by examining a number of salient parameters that influence their behavior. Variables such as the military value of atrocities, the power of the three parties, and the costs they face for war,
atrocities, and sanctions are all important. They help determine whether intervention will occur and if so at what threshold of atrocities, whether the state will observe the threshold or violate it, and whether the two sides can reach a bargain that avoids a war.

Our model shows that, as the critics contend, there are potential perverse outcomes associated with developing and strengthening an atrocities prevention policy. In particular, a credible antiatrocities regime strengthens weak actors and lowers the costs of war, which under certain conditions may make war and atrocities more likely. However, we also find that the potential negative effects of an antiatrocities regime are surmountable. In particular, by remaining neutral, by developing a multidimensional intervention repertoire, and by remaining sensitive to the balance of power, international actors can decrease the expected level of atrocities over time. On balance, the model shows that an investment in an antiatrocities regime need not be a road to hell. If proper care is taken along the lines we suggest, a stronger antiatrocities regime will have a net positive impact on preventing one of the globe’s most odious scourges: genocide and related forms of mass atrocities.

References


Rauchhaus, Robert W. 2009. “Principal-Agent Problems in Humanitarian Intervention: Moral Hazards, Adverse


**Supporting Information**

Additional Supporting Information may be found in the online version of this article at the publisher’s web site:

Complete Information Case

**Table 1:** Notation in the Game

Incomplete Information Case

**Table 2:** Equilibria with Incomplete Information