British Journal of Political Science
What the Enemy Knows: Common Knowledge and the Rationality of War
Thomas Chadefaux

Online supplementary material

This appendix adds a number of extensions to the simple game in the main text. In particular, I discuss in turn the effect of adding repeated interactions (finite and infinite), and of allowing players to communicate.

Repeated game (finite)

Here I describe a game in which there are two periods and an infinite number of types of B. I first describe the game formally, then show that there is no peaceful perfect equilibrium in this game. I then discuss in section the intuition for the proof, and finally show that the logic of this two-staged game applies to a game with an infinite number of stages.

The game

Setup Two states, A and B, negotiate over the partition of a territory X of size normalized to 1. At the beginning of the game, Nature chooses A's type $\phi \in \{w, s\}$, where w is selected with probability q and s with probability $1 - q$. It will sometimes be convenient to refer to a weak (strong) A as A^w (A^s).

Nature also chooses the signal θ that B receives about q—and hence about A's type.[33] θ is a random variable with probability mass function f, where f has support $[0, 1]$ and cumulative distribution $F(\cdot)$.[34] f is common knowledge, but the draw θ is only observed by B. In short, B has different possible levels of uncertainty about A's strength ϕ—in particular, the signal is perfectly informative if $\theta \in \{0, 1\}$ and is not if $\theta \in (0, 1)$—and A is uncertain about what B knows about A's strength.

Stage game After Nature's move, the game consists of $T = 2$ stages denoted by $t \in \{1, 2\}$. Each stage game follows a standard take-it-or-leave-it bargaining protocol with an outside option. Specifically, A first makes an offer $x_t \in [0, 1]$, where x_t denotes A's share of the pie. B can then: (a) accept, in which case the game ends with payoffs $x_t\delta^{t-1}$ for A and $(1 - x_t)\delta^{t-1}$ for B, where $\delta \in (0, 1)$ denotes a common discount factor; (b) reject, in which case the game proceeds to stage $t + 1$ (unless $t = T$, in which case 'war' ensues); or (c) fight (i.e., 'war'). In the event of a war, A wins the entire territory with probability p^ϕ, where $\phi \in \{w, s\}$, and B wins it with probability $1 - p^\phi$. Both players incur a cost of war c, and I assume that $p^s > p^w + 2c$.[35] I also assume for simplicity of exposition that $u_i(x) = x$ for all $x \in [0, 1]$ and $i \in \{A, B\}$.

Beliefs B has beliefs about A's type, which are affected both by Nature's signal θ and by A's sequence of offers $\{r_t\}$. More specifically, let $H_t = \{h_t\}$ be the set of all possible stage t histories. Then we define a belief function as a mapping $\mu^i : H_t \to [0, 1]$. Specifically, $\mu^B_{h_t} \equiv \mu^B(\phi = w|h_t) \in [0, 1]$ denotes B's belief that she is facing a weak A at time t, following history h_t; and $\mu^A_{h_t} \equiv \mu^A(\mu^B_{h_t}|h_t)$ denotes A's belief about B's belief—i.e., a mapping from each

[33] θ can be interpreted as: "B's prior is that A is weak with probability θ and strong with probability $1 - \theta$."
[34] Support over the entire interval is not necessary but facilitates presentation.
[35] Allowing for different costs of war for each player does not change the logic of the argument.

of B's beliefs about her own probability of being weak, onto a probability. Beliefs are updated according to Bayes' rule. As an example, initial beliefs immediately after Nature's choice of θ at time 0 are $\mu^B(\phi = w|\theta) = \theta$ and $\mu^A(\mu_t^B = x|\theta) = f(x)$.

Finally, define a 'peaceful' Perfect Bayesian Equilibrium (PBE) as an equilibrium in which war happens with probability zero. I now show that there is no peaceful PBE in this game.[36]

Analysis

Let r_t be the probability with which A^w makes a 'low' offer in stage t, i.e., an offer in the set $x_t^L = \{x_t : x_t \in p^w \pm c\}$, and $1 - r_t$ the probability with which she makes a 'high' offer instead, i.e., an offer in the set $x_t^H = \{x_t : x_t \in p^s \pm c\}$.[37]

Lemma 3. *In equilibrium, B accepts any offer $x_t^* \in x^L$.*

Proof (by contradiction). Suppose to the contrary that there is an offer $x_t^* \leq \max(x^L)$ that B rejects. Clearly this cannot be in the last round, as this would lead to war, with expected utility for B of $1 - p^\phi - c \leq 1 - p^w - c \leq 1 - \max(x^L) \leq 1 - x_t$. Similarly, since A can guarantee a payoff of x_t^* in the next round, B cannot credibly threaten to reject any $x_{t-1} \leq \max(x^L)$. □

It will be convenient to define θ^{r_t} as the implicit solution to

$$(1 - p^w - c)\mu^B_{\theta^{r_t},h_t} + (1 - p^s - c)(1 - \mu^B_{\theta^{r_t},h_t}) = x^H,$$

where $\mu^B_{\theta^{r_t},h_t} = \mu^B(\phi = w|\theta^{r_t}, h_t)$. In other words, θ^{r_t} is the value of the signal θ received by B for which, given history h_t, B is indifferent between accepting the high offer and fighting today.[38] This means that if B observes $\theta > \theta^{r_t}$ (i.e., B thinks A is likely to be weak), B rejects or fights upon receiving a high offer; for lower values of θ, B accepts.

I now first show that for an equilibrium to be peaceful, it must be that A makes a low offer with probability one in at least one of the two rounds.

Lemma 4. *In a peaceful equilibrium, either $r_1 = 1$ or $r_2 = 1$ or both.*

Proof (by contradiction). Suppose to the contrary that there is a peaceful equilibrium in which $r_1 < 1$ and $r_2 < 1$. Then with probability $(1 - r_1)(1 - r_2) > 0$, B receives high offers x^H in both rounds. Let $\tilde{\theta} = \max(\theta^{r_1}, \theta^{r_2})$ and note that B receives signal $\theta > \tilde{\theta}$ with probability $1 - F(\tilde{\theta}) > 0$, in which case B rejects any offer x^H. Therefore war occurs with probability $(1-r_1)(1-r_2)(1-F(\tilde{\theta})) > 0$, but this violates the assumption that the equilibrium is peaceful. □

[36]A much simpler proof of the absence of a peaceful PBE exploits the finite nature of the game, but unfortunately does not extend to the infinite version of this game. The proof goes as follows. First note that if $\theta = 1$, then B can guarantee an expected payoff of $1 - p^w - c$ by fighting. Since there is support for f on [0, 1], there exists a positive probability that A is facing a B with belief $\theta = 1$, i.e., that B is certain she is facing a weak A (I ignore some of the measure theoretic complications associated with continuous distribution—the result of course easily extend to the situation of a very large but finite number of type, each with a positive probability). Hence a peaceful PBE requires that A offer x_t such that B prefers accepting than fighting—i.e., A must offer at least $1 - p^w - c$ in t_1, or $(1 - p^w - c)\delta$ in t_2. But A cannot credibly promise to offer any x_2 such that $1 - x_2 > 1 - p^w - c$, as she would have an incentive to renege on that promise in t_2. Then the only way to prevent war with certainty is for A to offer x_1 such that $1 - x_1 = 1 - p^w - c$ with probability one—i.e., a separating equilibrium in round 1. But this setup is then the same as in the main text, and A prefers a pooling equilibrium, in which war happens with probability $f(\theta = 1)$. Unfortunately, this proof does not extend to the case of an infinitely repeated game. Indeed, with infinite periods, A could in fact credibly promise more than $1 - p^w - c$ in the next round (B could make sure of that by threatening punishment).

[37]It is easy to show that other offers are not viable as candidates for equilibrium.

[38]We assume that when B is indifferent between an offer and war, she accepts the offer. This is not a critical assumption.

The second step in the proof is to show that a strategy is an equilibrium strategy only if $r_t < 1$ for all t, and hence, per lemma 4, that there can be no peaceful equilibrium. I first show that $r_1 = 1$ cannot be part of an equilibrium strategy.[39]

Proposition 2. *Let $\tilde{\mathbf{r}} = (\tilde{r}_1 = 1, \tilde{r}_2)$ be A^w's strategy, where $\tilde{r}_2 \in [0, 1]$. Then $\tilde{\mathbf{r}}$ is not an equilibrium strategy.*

Proof (by contradiction). Suppose to the contrary that $\tilde{\mathbf{r}}$ is an equilibrium strategy, i.e., that A^w offers x^L with probability $r_1 = 1$. B accepts this offer (per lemma 3), leading at most to a payoff of $\max(x^L)$ for A^w, regardless of the value of $\tilde{r}_2$. Consider now a deviation to $\mathbf{r}' = (r_1', r_2')$, with $r_1' \in (\gamma, 1)$, where γ is the implicit solution to $F(\theta^\gamma) = \frac{p^w+c}{p^s-c}$ and $r_2' \in [0, 1]$.[40] In that case, B receives an offer x^L with probability r_1', which she accepts, and an offer x^H with probability $1 - r_1'$. A expects B to accept x^H with probability $\mu_1^A = \mu^A(\mu_1^B \geq \theta^{r_1'})$. If B rejects it, then A can always ensure her minimum payoff of $\min(x^L) = p^w - c$ in the next round, regardless of beliefs. In sum, A^w's expected payoff from deviating to r_1' is therefore at least:

$$r_1' x^L + (1 - r_1') \left[\delta \min(x^L)\mu_1^A + \min(x^H)\left(1 - \mu_1^A\right)\right],$$

which is greater than x^L for all δ if

$$\mu_1^A < 1 - \frac{p^w + c}{p^s - c}, \tag{1}$$

and note that

$$\mu_1^A = \mu^A(\mu_1^B \geq \theta^{r_1'}) = \int_{\theta^{r_1'}}^{1} f(\theta) d\theta = 1 - F(\theta^{r_1'}).$$

Furthermore, I assumed $r_1' > \gamma$, so $\theta^{r_1'} > \theta^\gamma$.[41] Therefore, $1 - F(\theta^{r_1'}) < 1 - F(\theta^\gamma)$. Finally, γ was defined as the implicit solution to $F(\theta^\gamma) = \frac{p^w+c}{p^s-c}$, so

$$\mu_1^A = 1 - F(\theta^{r_1'}) < 1 - F(\theta^\gamma) = 1 - \frac{p^w + c}{p^s - c}.$$

This means that (1) holds and hence that r_1' is a profitable deviation from $\tilde{r}_1 = 1$. But this contradicts the assumption that $\tilde{\mathbf{r}}$ is an equilibrium strategy. □

I now show that $r_2 = 1$ is also not an equilibrium strategy.[42] The first step in this part of the proof is to show that $r_2 = 1$ can only be part of an equilibrium strategy if r_1 was high enough to screen out the low types of B prior to stage 2.

[39]The logic is the same as the model in the text: A^w can improve upon a separating equilibrium by offering a dose of pooling such that extracts the maximum from Bs who might think she is strong.

[40]Recall that $F(\cdot)$ is the cumulative distribution function of f.

[41]Intuitively: if r_1' is larger, then the threshold at which B rejects is higher, i.e., it takes a larger level of certainty for B to reject.

[42]The intuition for this part of the proof is that for A to make separating offers in stage 2, she must be confident enough that B thinks she is weak. Otherwise she would prefer a dose of pooling, as in the first stage. But in order to be confident enough, it must be that r_1 (in the previous stage) was high enough that most types of B have been screened out (i.e., those for whom θ is small).

Lemma 5. *Suppose* $\mathbf{r}^* = (r_1^*, r_2^* = 1)$ *is an equilibrium strategy. Then* r_1^* *is an implicit solution to* $\mu(\mu_2^B > \theta^{1-\varepsilon}|h_2, r_1^*) \geq 1 - \frac{2c}{p^s - p^w}$ *for all* $\varepsilon \in (0, 1]$.[43]

Proof (by contradiction). Suppose to the contrary that $\mu(\mu_2^B > \theta^{1-\varepsilon}|h_2, r_1^*) < 1 - \frac{2c}{p^s - p^w}$ for some arbitrary ε. I now show that A can profitably deviate from $r_2^* = 1$ and hence that $\mathbf{r}^*$ cannot be an equilibrium strategy. First note that A's expected utility from $r_2^* = 1$ in stage 2 following any history is at most $p^w + c$. Consider now a deviation to $r_2' < 1 - \varepsilon$. Then, with probability r_2', A offers $p^w + c$ and with probability $1 - r_2'$ she offers $p^s - c$, in which case B rejects if $\mu_2^B \equiv \mu_2^B(\phi = w|h_2) \geq \theta^{1-\varepsilon}$, giving A a payoff of at least $p^w - c$. So A's expected value for offering r_2' is at least

$$u(r_2') = r_2'(p^w + c) + (1 - r_2')\left[(p^w - c)\mu_2^A + (1 - \mu_2^A)(p^s - c)\right],$$

with $\mu_2^A = \mu(\mu_2^B \geq \theta^{1-\varepsilon}|h_2, r_1^*)$.

Since $\mathbf{r}^*$ is an equilibrium strategy, as assumed, then it must be that $u(r_2^*) \geq u(r_1')$ for all r', and hence that $\mu_2^A \geq 1 - \frac{2c}{p^s - p^w}$, which contradicts our initial assumption. □

I now show that there exists a profitable deviation from $\mathbf{r}^*$.

Proposition 3. *Let* $\mathbf{r}^* = (r_1^*, r_2^* = 1)$*, with* r_1^* *as in lemma 5. Then* $\mathbf{r}^*$ *is not an equilibrium strategy if* $F(\theta^{r_1^*}) > \frac{1}{2}$.

Proof. To see this, note first that A's expected value from strategy $\mathbf{r}$ is:

$$E[u_A(\mathbf{r})] = r_1(p^w + c) + (1 - r_1)\int_0^1 f(\theta)V(x_1^H|\theta, \mathbf{r})d\theta, \tag{2}$$

where $V(x^H|\theta, \mathbf{r})$ is A's continuation value for offering x_1^H to a B who observed signal θ and given strategy $\mathbf{r}$. Consider first strategy $\mathbf{r}^*$ and note that

$$V(x^H|\theta, \mathbf{r}^*) = \begin{cases} p^s - c & \text{if } \theta \in [0, \theta^{r_1^*}), \\ \delta(p^w + c) & \text{if } \theta \in [\theta^{r_1^*}, 1] \end{cases}$$

Suppose now that A deviates to $\mathbf{r}' = \{r_1', r_2'\}$, where $r_1' = r_1^* - \varepsilon$ with $\varepsilon > 0$, and $r_2' = 1 - \gamma$, with $\gamma > \frac{1 - r_1^*}{\delta(1 - r_1^* + \varepsilon)}$. Then:

$$V(x^H|\theta, \mathbf{r}') = \begin{cases} p^s - c & \text{if } \theta \in [0, \theta^{r_1'}), \\ \delta(r_2(p^w + c) + (1 - r_2)(p^s - c)) & \text{if } \theta \in [\theta^{r_1'}, \theta^{r_1^*}), \\ \delta(r_2(p^w + c) + (1 - r_2)(p^w - c)) & \text{if } \theta \in [\theta^{r_1^*}, 1] \end{cases}$$

Intuition: those with $\theta \geq \theta^{r_1^*}$ still reject, as before, because they are even more convinced to be facing a weak A. But those with θ a little below θ^* also now reject, but in the second round face a different offer.

[43]Intuition: r_1^* is the smallest value of r_1 such that enough Bs will be screened out (i.e., A's posterior probability of facing B with beliefs $\mu_2^B > \theta^{r_2'}$ is at least $1 - \frac{2c}{p^s - p^w}$. I.e., A is then sufficiently confident that B thinks she is weak, and hence such that there is no profitable deviation from the separating equilibrium.

Now we just need to show that $E[u_A(\mathbf{r}')] > E[u_A(\mathbf{r}^*)]$, which can for simplicity be rewritten as the sum of conditional expectations for all three possible sets of values of θ:

$$\int_0^1 E[u_A(\mathbf{r}') - u_A(\mathbf{r}^*)|\theta]d\theta > 0$$

Using (2), algebra yields:

$$E[u_A(\mathbf{r}') - u_A(\mathbf{r}^*)|\theta] = \begin{cases} (x^H - x^L)\varepsilon & \text{if } \theta < \theta^{r_1'} \\ x^L\varepsilon & \text{if } \theta^{r_1'} < \theta < \theta^{r_1^*} \\ (\delta x^L - x^H)\varepsilon & \text{if } \theta > \theta^{r_1^*} \end{cases}$$

Putting these together:

$$E[u_A(\mathbf{r}')] - E[u_A(\mathbf{r}^*)] = (x^H - x^L)\varepsilon \int_0^{\theta'} f(\theta)\, d\theta + x^L\varepsilon \int_{\theta^{r_1'}}^{\theta^{r_1^*}} f(\theta)\, d\theta - (x^H - \delta x^L)\varepsilon \int_{\theta^{r_1^*}}^{1} f(\theta)\, d\theta$$

Rearranging, we find $E[u_A(\mathbf{r}')] - E[u_A(\mathbf{r}^*)] > 0$ if:

$$\int_{\theta^{r_1^*}}^{1} f(\theta)\, d\theta < \frac{(x^H - x^L)\int_0^{\theta_{r_1'}} f(\theta)\, d\theta + x^L \int_{\theta^{r_1'}}^{\theta^{r_1^*}} f(\theta)\, d\theta}{x^H - \delta x^L} \tag{3}$$

Note that $\int_0^{r_1'} f(\theta)d\theta$ can easily be shown to be smaller than the right hand side of inequality (3), and so A has an incentive to deviate if $1 - F(\theta^{r_1^*}) < F(\theta^{r_1'})$. Moreover, $r_1' = r^* - \varepsilon$ can be made arbitrarily close to r_1^*, and hence A will have an incentive to deviate if $F(\theta^{r_1^*}) > \frac{1}{2}$ □

Theorem .1. *Suppose $F(\theta^{r_1^*}) > \frac{1}{2}$, where r_1^* is as in lemma 5. Then there is no Peaceful Perfect Bayesian Equilibrium in this game.*

Proof. The proof follows immediately. Lemma 4 shows that in a peaceful equilibrium, either $r_1 = 1$ or $r_2 = 1$, or both. Yet propositions 2 and 3 show that any strategy that offers $r_t = 1$ in any t cannot be an equilibrium strategy. □

Intuition

In this section, I discuss the intuition for why adding periods is insufficient to avoid war. I focus on a simplified version of the model discussed in the section above.

Suppose that there are two periods and three types of B. For ease of exposition, call them the 'skeptical' type (i.e., B is certain or quite certain that A is weak), the 'medium' type (e.g., B thinks there is a 50% chance A is weak), and the 'gullible' type (B is quite convinced that A is strong). Suppose furthermore that the distribution of B types is such that the skeptical type is *a priori* sufficiently unlikely.

First, note that if there is to be a peaceful equilibrium, A must adopt a separating strategy in at least one of the two periods (i.e., a strategy whereby a weak (strong) A offers a low (high) offer). Otherwise (i.e., if A includes some pooling in both rounds), there is a positive probability that a skeptical B would receive two high offers from A, and reject both, leading to war. Second, note that this separating equilibrium cannot take place in the first stage, for the same reasons as in the one-stage game: because all three types remain in the first stage, A is better off incorporating at least some pooling in her strategy in order to extract as much as possible from the 'gullible' Bs.

So if there are to be separating offers, they must be made in the second stage. But in the second stage, it is only rational for A to adopt a separating strategy if she thinks the 'gullible' Bs have already been weeded out. Indeed, if all types of B are as likely in the second stage as they were in the first stage, then we are again faced with the same situation as in the text, and pooling is rational in order to extract as much as possible from gullible Bs. If, instead, all but the 'skeptical' types of B have been weeded out, then pooling would be too risky a strategy and A has an incentive to make separating offers.

The existence of a peaceful equilibrium, then, hinges on how many types are left in the second round. If both the skeptical and the medium types are left, then we are back to the situation in the main text, and A will be tempted to introduce an element of pooling—and risk war with the skeptical types. War will therefore be avoided for sure only if both the gullible type *and* the medium type have been weeded out by A's first round offer, leaving only the skeptical type in the second round. In that case only will A offer make a low offer with probability one.

Unfortunately, A could instead only weed out the gullible type in the first stage, keeping both the medium and 'skeptical' types in the second round. If so, then she should incorporate an element of pooling in the second round, leading the skeptical types to receive a high offer with some probability, which they will reject, leading to war.

The decision of whether to weed out both the gullible and the medium type in the first round, or only the gullible type, depends on the relative probability of facing each type. In particular, if the gullible type and the medium type are sufficiently likely, then A will be tempted to ignore the skeptical type, and hence to first weed out the gullible type, and in the second round to weed out the medium type, leaving the skeptical type to fight.

Repeated game (Infinite periods)

The reader might think that an infinite number of periods would solve the problem, since each type could be weeded out in each round, over an infinite number of rounds. But this is in fact not the case. To understand why, note that there are limits to how long a skeptical B would wait for a satisfying offer. Even if A^w could somehow credibly promise to make an offer $p^w - c$ in time t^*—where t^* is the time it takes for A to screen the 'gullible' types of B—B may not be willing to wait this long.

More precisely, note that B can fight today with expected payoff $1 - p^w - c$. So B will only prefer an offer $1 - p^w + c$ (the most A can credibly offer, regardless of future punishments) in time t^* to war if:

$$
\begin{aligned}
\delta^{t^*}(1 - p^w + c) &> 1 - p^w - c \\
\Rightarrow t^* &< \log_\delta\left(\frac{1 - p^w - c}{1 - p^w + c}\right)
\end{aligned}
$$

t^* is an upper bound on the duration of the negotiation, and as a result on the number of possible types that a peaceful equilibrium can possibly accommodate. Note that t^* may in fact be very small. As an example, with $p^w = 0.4$, $c = 0.1$ and $\delta = 0.9$, $t^* \lesssim 3$, so that 4 possible types of B may be enough to prevent a peaceful PBE.

Suppose now that the game in the main text is modified to allow B to send A a message m from a set M of possible messages immediately after Nature's move. After A receives m, the game proceeds exactly as in the original game, with the same payoffs.

Let $\phi^B \in \{n, nk\}$ refer to B's two different types. Furthermore, for all $m \in M$ and ϕ^B, let $f(m, \phi^B)$ be the probability that type ϕ^B announces message m in a given equilibrium. I refer to $f(\cdot)$ as B's message strategy. Upon observing m and given strategy f, A updates her beliefs to $\mu_A(\phi^B = \phi|m, f)$ and responds with a demand $x(m, f)$, which associates an offer x to every possible message m.

It will be convenient to define r^* as the equilibrium probability with which, in the game without communication, A makes an offer $x \in x^L$, where $x^L = \{x : x \in p^w \pm c\}$.

Proposition 4. *Suppose that war happens with positive probability in the game without communication. Then war happens with the same probability in the game with communication.*

Proof (Sketch). Suppose that there exists a message strategy $f^*(m, \phi^B)$ such that war happens with a lower probability than in the game without communication. Since war happens in the original game when a knowledgeable B receives an offer $x \notin x^L$, the probability of war in equilibrium can only be reduced if $Pr(x(m, f^*) \in x^L) > Pr(x \in x^L)$, i.e., if observing m leads A to make an offer $x \in x^L$ with probability $r' > r^*$.

But note that this is rational for A only if m is such that $\mu_A(\phi^B = k|m, f^*) > \mu_A(\phi^B = k)$, i.e., if A's belief that she is facing a knowledgeable type increases as a result of receiving message m and given strategy f^*.

If such a message exists, however, then clearly a non-knowledgeable type of B would also want to send it. But then it must be that $\mu_A(\phi^B = k|m, f^*) = \mu_A(\phi^B = k)$, which contradicts the result above. Therefore there can be no informative signal that reduces the probability of war, and hence war happens with the same probability in the game with communication. □