Noisy warning systems and nuclear war*

James D. Fearon
Stanford University

July 21, 2025

Abstract

Thomas Schelling (1960) argued that mutual fears of a violent preemptive attack could spiral, leading to an actual attack even when both sides preferred no attack. His main concern was nuclear war, but it seems relevant in other settings, including police shootings and military coups. I analyze a model of a coordination problem in which players observe noisy signals of whether they are already under attack. How the players interpret signals proves to be endogenous to equilibrium expectations and is strongly influenced by the size of the first-strike advantage. With small advantage, as is plausible for the nuclear case under mutual assured destruction, the risk of inadvertent escalation is very low for any degree of accuracy of the warning system, and players set the probability of a false positive to close to zero and the probability of a false negative close to one. With larger first-strike advantages, tragic escalation is unlikely for very accurate *or* very noisy warning systems, and highest for "so-so" signal accuracy. In these cases the choice of "hair trigger-ness" is a case of strategic complements. Players are driven by the negative externality of the other side's choice to be jumpier than they would want to be if they could commit to be more relaxed. The results on endogenous interpretation of signals may help explain why surprise attacks are routinely followed by discovery of what appear after the fact to be obvious warning signs. It also shows how pursuit of counterforce nuclear capability can actually reduce a state's resolve and payoffs in a crisis by increasing crisis instability.

*An earlier version was presented at the 2019 Annual Meetings of the American Political Science Association, Washington DC, August 29-September 1. Incomplete draft, comments welcome.

> If I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is danger of an outcome that neither of us desires. Even if he prefers just to leave quietly, and I wish him to, there is danger that he may *think* that I want to shoot, and shoot first. Worse, there is a danger that he may think that *I* think that *he* wants to shoot. Or he may think that *I* think that *he* thinks that *I* want to shoot. And so on.
>
> Thomas Schelling, "The reciprocal fear of surprise attack" (1960)

1 Introduction

What are the most likely paths by which states could end up using nuclear weapons? Although thankfully still thin, accumulating historical evidence suggests two main possibilities. First, leaders or their agents in the military receive indications and warning that makes them think that the other side has already begun, or is about to begin, an attack that could cripple their nuclear capabilities and more. This makes them want to launch to avoid the disadvantages of being the victim of a first strike. Second, a nuclear-armed state is losing badly in a conventional fight. It decides to detonate a nuclear weapon to signal willingness to do more, or risk massive escalation by the first pathway, to pressure the other to back off.[1]

This paper investigates the strategic logic of the first pathway, which was first clearly characterized and analyzed by Thomas Schelling in the article from which the epigraph is taken. The burglar story illustrates what he and others in the 1950s feared was a plausible path to nuclear war. Schelling thought that if decision-makers in the US and Soviet Union saw an advantage to launching their nuclear weapons first rather than second – a "first-strike advantage" – then there could be strategic dynamics that would generate a nuclear war even if it was known that all preferred no

[1] A third possibility is that a weapon could be launched due to a pure accident or command and control failure, or conceivably by a "crazy" or internally dysfunctional leadership, unrelated to fear of possible attack in progress by an adversary. This possibility can't be eliminated and it will play a role in the arguments that follow, as it is a "seed" for fears about deliberate use in response to alarm signals. Sagan (1993) provides evidence on numerous accidents in nuclear-weapons command and control systems during the Cuban Missile Crisis and in the 1968 crash of a US bomber carrying four thermonuclear bombs. The latter case and a number of his examples from the missile crisis are in this class of pure accidents. Many others are accidents in the sense of false positives in the warning systems that could trigger deliberate launch, which is the focus of this paper.

war to any nuclear war (first strike or not).

Several incidents in nuclear history suggest the concern is not groundless. During the Cuban Missile Crisis, on October 27, 1962, U.S. Navy destroyers began dropping signaling depth charges in an attempt to force a Soviet submarine armed with nuclear torpedoes (unknown to the U.S.!) to surface. Out of contact with the Russian Navy and experiencing low oxygen and temperatures greater than 115°, the submarine's captain and the political officer feared that war had already begun (they did not know these were signals) and wanted to fire the torpedoes. They were opposed by Chief of Staff Vasily Arkhipov, whose assent was also required, and who argued successfully that they should surface. Had the nuclear torpedoes been used the likelihood of escalation to large-scale nuclear war by both countries was probably very high (Savranskaya, 2005).[2]

On September 26, 1983, in a political and military context where Soviet leaders increasingly feared a NATO attack and possible nuclear first-strike by the U.S., the Soviet early warning system Oko mistakenly reported launch of five ICBMs towards Russia. Stanislav Petrov, the relevant duty officer, had suspicions about the warning and chose not to report it, against orders. What leadership would have done if given the report and told they had minutes to decide on a response is not known.

Much more recently, on January 13, 2018, a combination of miscommunication and human error led to Hawaii's state government issuing an official alert urging residents to immediately seek shelter because of an incoming ballistic missile. Not coincidentally, this was in the context of escalating tension between the Trump administration and North Korea over the latter's nuclear program and tests. Many indications suggest that the Trump administration was seriously contemplating a first-strike (Jackson, 2019, chap. 7). News of the alert in Hawaii, coupled with other

[2]There were four Soviet submarines armed with nuclear torpedoes in the group, and this was not the only encounter that could easily have led to a one being fired. Savranskaya (2005, 242) writes "In interviews and in memoirs, all the Soviet captains recalled their state of extreme tension and confusion in a situation where the war above could have begun any time while they were trying to evade their pursuers in the submerged position, with no communication with the outside world. The captains also anticipated that in the situation where the nuclear exchange either became inevitable or had already begun, Moscow would want them to use their special weapons first, as they were instructed before departure." See Sagan (1993) for his discovery of multiple dangerous cases of false warning of attack in US systems at the height of the Cuban Missile Crisis.

possible random indications they were receiving, could have led the North Korean government to assume that an attack had begun or was about to begin. Multiple other events and incidents around this time and in the previous six months – mainly missile launches by North Korea and U.S. force movements and exercises in the region – could also have produced incorrect inferences about impending attacks or attacks in progress.[3]

Inadvertent escalation due to noisy warning systems and "reciprocal fears" concerns is plausibly linked to tragic outcomes in other settings besides interstate nuclear crises. Philip Roessler (2011, 2017) argues that civil wars have often emerged as a result of a mutual fears-driven falling out among former co-conspirators who had successfully taken power themselves. He summarizes:

> elite accommodation in the shadow of the coup d'état can give rise to an internal security dilemma, as power holders, fearful that the other side is going to violate its commitment to sharing power, maneuver to defend their privileged positions. But such action merely increases uncertainty and intrigue that factions in the regime are plotting to seize or consolidate power on their own at the expense of other stakeholders. Rising mutual fears lead allies-turned-rivals to adopt more extreme measures to defend themselves until eventually a point of no return is reached and both sides become convinced that they will be eliminated in the future (Roessler, 2011, 211).

Similarly, when police encounter a suspect whom they think may be armed and dangerous, they sometimes need to make split-second judgements based on perceptions of the situation and the person's actions (they receive noisy signals). Aware of this, suspects are themselves often terrified of being shot, which does not necessarily make for calm reactions, which may then feed back on a police officer's propensity to shoot.[4]

At the heart of Schelling's analysis of the "reciprocal fears" problem is a Stag Hunt game, a much-studied coordination problem of wide applicability in political, economic, and social situations. A 2×2 example is shown in Figure 1. The game has two pure-strategy Nash equilibria, (*Don't Attack, Don't Attack*) and (*Attack, Attack*). Even though both players are better off if they

[3] Jackson (2019).

[4] "Hands up!" is a time-honored convention intended to produce a clear signal. Research finds that police quite often shoot suspects who are beginning or have just made this gesture, possibly because decision time to preempt firing by the suspect is very short (1/3 second on average in this experimental study) (Aveni, 2008).

Figure 1: Stag Hunt ($1 > \alpha > 0 > -\beta$)

		Player 2: *Don't attack*	Player 2: *Attack*
Player 1	*Don't Attack*	$1, 1$	$-\beta, \alpha$
	Attack	$\alpha, -\beta$	$0, 0$

coordinate on the former, attacking arguably has a rational appeal if α is close to 1 and β is large. In such cases, (*Attack, Attack*) is said to be "risk dominant," an equilibrium-selection principal in competition with the Pareto-dominance principle that recommends (*Don't Attack, Don't Attack*).[5]

Schelling tried several ways to formalize a reciprocal-fears strategic dynamic, using game-theoretic tools available at the time. He found it difficult to produce a model based on natural assumptions that generates a dynamic escalation of beliefs leading to attack. He saw the most promise in a version that tried to incorporate fallible *warning systems*. In that model, decision-makers implicitly choose whether to attack on the basis of unmodelled noisy signals of whether the other side has already begun an attack. The strategically interesting and important choice then shifts to how sensitive to make one's warning system. Schelling's discussion suggests that he thought warning-system sensitivity choice could be a case of strategic complements: The more you are on a hair trigger, the more it makes sense for me to be on a hair trigger, to our mutual detriment in terms of risk of conflict from a false positive (e.g., mistaking a seagull for a missile, or a cell phone for a gun in the case of police shootings).

This paper explicitly models noisy signals of action choices in a Stag-Hunt-type of coordination problem. Players observe a signal of whether an attack is already in progress when they make their decision about whether to attack. There is some chance, which can be small, that each player either attacks due to an accident, or is a "crazy type" who attacks no matter what.

I focus on how three factors interact to render a situation more or less prone to inadvertent escalation:

[5]Formally, both attacking risk dominates neither attacking in the game of Figure 1 when $\alpha+\beta \geq 1$. This condition implies that a player who thinks the other other is equally likely to attack or not attack prefers to attack.

1. The size of the first-strike advantage and the second-strike disadvantage (payoffs α and β);
2. the players' prior beliefs about how likely it is that the other would attack no matter what, or might attack accidentally; and
3. the precision of the warning technology, meaning how accurate or noisy are the signals.

Unsurprisingly, greater first-strike advantage and belief in greater probability of "crazy" types or accidents makes inadvertent escalation more likely. The effects of signal noise, by contrast, are subtle and complicated. It is also instructive to learn how, in this model, players' interpretation of signals is a function of prior beliefs that are themselves endogenous to equilibrium expectations. I find that:

1. Tragic escalation is least likely when signals are either very precise *or very noisy.* Escalation is most likely for "so-so," or intermediate warning systems. The main reason is that players will tend to rationally discount warning from noisy technologies, requiring a very strong and clear signal to attack. And because each expects the other to discount heavily, their endogenous *prior beliefs* about an attack are low, which reinforces or justifies heavily discounting signals.

   With intermediate signal precision and relatively large first-strike advantages, the players in effect face a Prisoners' Dilemma-like problem: Players do not fully internalize the "negative externality" on the other player from making their own warning system more "hair trigger." Tragic escalation can be more likely and the players worse off than if they could commit to ignore signals or have no warning system at all. In extreme cases, introduction of a "so-so" warning technology can actually eliminate all but the bad equilibrium in which both players attack regardless of signal.

2. Discounting of signals is greater the smaller the first-strike advantage. Indeed, for small first-strike advantages, there is virtually no chance of inefficient escalation for *any* signaling system precision.

   Thus in the case of nuclear war, where we would usually expect the first-strike advantage to be seen as small relative to the difference between peace and any nuclear exchange, the results imply that alarms will tend to be heavily discounted and monitors quite skeptical of all but multiple, very clear signals.

   By contrast, in the case of police encounters and coup fears, where first-strike advantages are plausibly much larger on average, in equilibrium players endogenously apply more of a "hair trigger." The likelihood of unwanted escalation is much higher, and highest for "so-so" signal precision.

3. With small first-strike advantages, in equilibrium players set the probability of a false positive to be very small, and the probability of a false negative to be close to one! That is, they deliberately heavily discount warnings, so that the probability that they mistakenly start a war (attack) is very small. But at the same time the probability that they will incorrectly dismiss warning of an actual attack in progress is *nearly certain*.

   This observation could help explain why, after famous surprise attacks, analysts always find what appear to them to be clear signs of impending attack that were wrongly dismissed or missed. Ex ante, it can make sense to do so to avoid false positives in situations where the prior belief that other wants to attack is low (which is implied by "surprise" attack).

The next section discusses related, mainly game-theoretic literature; it can be skipped if one is not interested in differences between the model here and other approaches to the Stag Hunt problem. The third section presents the model. The fourth considers the extreme or polar cases of perfect signals and no signals, following by results for the model with noisy signals in section 5. Section 6 briefly discusses an implication of the results for understanding intelligence failures and surprise attacks. Section 7 concludes with caveats and possible extensions.

2 Related literature

The literature on Stag-Hunt-like coordination problems is extensive, with much focus on different approaches to equilibrium selection. These include evolutionary models, global games, and models that introduce large variation in payoff types.[6] In evolutionary models, fixed rules revise strategies each period as a function of prior outcomes, typically (in coordination games) leading to convergence on one or another equilibrium depending on initial conditions and chance. Under some conditions risk dominant equilibria are selected in Stag Hunt games. In global games, players observe noisy signals of the game's true payoffs. This again can, under certain assumptions, yield selection of risk dominant equilibria. A number of studies have used the global games approach to examine political conflict situations, such as protests, revolutions, coups, and elections (e.g., Boix and Svolik, 2013;

[6]The evolutionary and global games literatures are vast. Some influential examples include Ellison (1993); Kandori, Mailath and Rob (1993); Carlsson and Van Damme (1993); Morris and Shin (2003); Angeletos, Hellwig and Pavan (2007); Chassang (2010). Baliga and Sjöström (2004) enrich the type space in Stag-Hunt problems in a way that can lead to what Schelling (also Kuran 1991) called "tipping" and for certain type distributions implies a unique equilibrium (see also Baliga, Lucca and Sjöström, 2011; Baliga and Sjöström, 2012).

Dewan and Myatt, 2008; Edmond, 2013; Little et al., 2012; Little, 2016, 2017).

The approach in this paper is different. Whereas in global games players get a noisy signal of their own and others' *payoffs* for different outcomes, here players get a signal of the other player's *action*, which allows them to update about whether the other has already started an attack. This is arguably a realistic approach for a large range of strategic situations where there is some relatively decisive, discrete action choice at issue – attacking, shooting, starting a coup or purge, initiating a nuclear weapons program, going to protest, building a factory, entering a market, proceeding at a 4-way intersection, stepping through the door, stepping left or right on the sidewalk, and many more. Players can usually observe small physical or other signals of what the other is about to do on the big choice.

In other words, in the real world people encounter coordination problems in continuous time, and it often happens that taking the major action at issue requires some preparatory or initiating actions that are imperfectly observable or interpretable. The model examined here is not literally in continuous time, but approximates this with a more tractable representation where one player chooses first but the players don't know who is the first mover (following Powell 1989). This set up allows me to augment the game with noisy signals about whether an attack is already occurring.

Powell (1989) and Baliga and Sjöström (2004) are the only explicit attempts to model Schelling's reciprocal fear of surprise attack problem that I can find.[7] In Powell's game, two states are assumed to be in a nuclear crisis, and they are unsure if the other has already chosen to attack when they make their choice. He shows that if states' preferences are such that they want to attack only if they estimate that the chance they are themselves under attack is at least 50%, the game has no equilibrium with any attacks. The reason is that the game structure implies that at least one state can obtain the "concede" payoff with at least .5 probability by quitting, so that this is always better than any nuclear war payoff (under the MAD assumption that nuclear war is worse than ceding the issue even if you strike first).

[7]Bracken and Shubik (1993) consider two small extensions to Powell's game.

The model here differs in that (a) not attacking does *not* assure that the first-mover can end the crisis (here, avoid war) for sure, because the other may be crazy or have an accident, or not-crazy but still get a high signal and decide to attack; (b) there is always positive probability of an attack, due to aggressive types or accidents; and (c) there are noisy signals of the other's action.[8]

Noisy signals are arguably a precondition for representing the strategic problem Schelling described in the vignette. Because the model here is in effect "one shot" (no pun intended), it does not allow for a back-and-forth escalation of beliefs as a result of a long sequence of misread signals. However it does produce the possibility of "one round" of tragic escalation due to a reciprocal fear of surprise attack. This occurs when a randomly high enough signal causes a nervous state to initiate attack even though the other has not begun to attack.

Baliga and Sjöström (2004) argue that the reciprocal fear of surprise attack logic is captured formally by a Stag Hunt game with a distribution of payoff types that produces "tipping" to a unique equilibrium. That is, knowing that there are a few types who would attack no matter what, we know also that there are some types who want to attack given that the dominant strategy types will attack, and so on, leading all types to attack. Schelling certainly did think this kind of dynamic was important in the real world, as shown by his work on tipping games with many players (Schelling, 2006).[9] However there is no sign in the reciprocal fear of surprise attack essay that I can see that suggests he saw this logic as relevant there. Perhaps he did, and even if he didn't it is of course still a possible interpretation or explanation. Suffice to say that the approach here focuses not on inferences from common knowledge of continuous distributions, but instead on inferences from noisy signals of actions.

Finally, two recent articles by Acemoglu, Wolitzky and Jackson have developed dynamic

[8]In Powell's model, quitting the crisis is rather like "Hands up!", a signal that is intended to be an unambiguous commitment not to attack. In the present model, there is no way to guarantee that the other state does not see a signal that could indicate attack.

[9]However, he does not appear to have imagined that tipping as equilibrium selection could or was likely to operate by introspection based on prior beliefs about distributions of types. Rather, he clearly had in mind a more evolutionary or adaptive process wherein individuals observe last period's actions and then make a new choice.

models of Stag Hunt games with overlapping generations of players, in order to study ways that a society might transition between good and bad equilibria (Acemoglu and Wolitzky, 2014; Acemoglu and Jackson, 2014). Players in one generation observe sometimes noisy or incomplete signals of past play, which can make for inefficient escalation over the course of generations of players.

3 Model

Two states, 1 and 2, will decide whether to attack the other after observing signals. The game form is as follows.

1. Nature chooses whether state 1 or state 2 moves first, with equal probability. Let $i \in \{1, 2\}$ be the first mover and $j \neq i$ be the second mover. The players *do not observe* whether they are first- or second-mover (as in Powell 1989).

   I will sometimes use $k \in \{1, 2\}$ and $-k$ to index a generic state and its adversary, since they do not know if they are first or second mover.

2. Nature also chooses the states' types, with probability $\epsilon \in (0, 1)$ that each state is a crazy type who has preferences that make attacking a dominant strategy.[10]

   States know their own type but not the other state's type.

3. Nature sends a signal to the first mover, state i. This is a number z_i drawn from the cumulative distribution function F_0.

4. If state i is not an aggressive type, i chooses whether to attack: $a_i \in \{0, 1\}$.

5. Nature sends a signal to the second mover, state j, a number z_j drawn from the cumulative distribution function F_0 if i did not attack, and from F_1 if i did attack. F_1 likelihood-ratio dominates F_0, which means that higher signals are (weakly) more likely if j is under attack, for all signals that have positive probability.[11]

6. If state j is not an aggressive type then j chooses whether to attack: $a_j \in \{0, 1\}$.

[10] Alternatively, with some minor modifications to payoff expressions, ϵ can be the probability that each state attacks by accident or mistake when making its choice. For simplicity I will keep to the "aggressive type" version in what follows.

[11] Note also that "likelihood ratio dominates" implies that F_1 first-order stochastic dominates F_0, meaning that $F_1(z) < F_0(z)$ for all z.

To summarize: The players do not know whether the other player has already made her choice to attack or not, but each observes a signal of whether an attack is in progress when making her own decision. Each player knows that there is some chance that other might be a type that just wants to attack no matter what.

Payoffs are as in the Stag Hunt of Figure 1. That is, for non-crazy types if neither attacks, they get $(1, 1)$. They get $(0, 0)$ if both attack, and $(-\beta, \alpha)$ or $(\alpha, -\beta)$ if just one attacks.

A complete strategy for a state is a choice of whether to attack (or a probability of attacking) for all possible signals.

For the signal distributions, in the main case considered below I let $F_a(z)$, for $a \in \{0, 1\}$, be the Normal distribution with mean a and variance σ^2. In this case the likelihood ratio for attack given signal z, $L(z) = f_1(z)/f_0(z)$, is increasing without bound, so that higher signals increase one's belief that one is under attack. Further, it will be shown that in this case there is always positive probability of a signal strong enough that a state would definitely want to attack. The variance term σ^2 indicates how informative the signal is, or in other words how good is the warning system. For small variance (good warning system), signals are highly diagnostic on average, whereas with high variance (bad warning system), $L(z)$ is usually closer to 1, so that a state can't update much based on the signal.[12]

Why have crazy types or risk of accidents in the model? Suppose instead that there were no crazy types (or no accident risk), so $\epsilon = 0$ rather than $\epsilon > 0$. Then the game always has a perfectly peaceful equilibrium in which neither state attacks regardless of the signal it gets. If the players are completely confident that the other will not attack, they can and should completely ignore the signals.

For most settings (nuclear conflict, police encounters, possible coups) it is more realistic to think that the players cannot be completely confident. Accidents or crazy-type risk are realistic

[12]The Normal distribution is convenient but, with the addition of some boundary conditions, most results will hold for any distribution that has a strictly increasing likelihood ratio. This includes all exponential family distributions. Note that setting the means at zero and one is without loss of generality.

ways of undermining perfect confidence in the model. With $\epsilon > 0$ and the continuous signal distributions just described, it will follow that the game does not have an equilibrium with zero probability of attacks by non-crazy types. This is because there is positive probability (though possibly vanishingly small) that a state happens to get a very high signal, which convinces it that attack is almost certainly occurring and so attacking in reply is the safest course. Substantively, we are considering situations where there is always some signal that could convince a player that it should attack to avoid getting the second-strike payoff $-\beta$.

4 No signals and perfect signals

It is useful to begin by analyzing two polar cases for the signaling technology: no signals and perfectly informative signals.

If the states observe no signal, or equivalently, a perfectly uninformative signal ($F_0 = F_1$), then we have[13]

Proposition 1. *With no signals, the game always has a pure strategy equilibrium in which both states attack for sure. If $\epsilon < \underline{\epsilon} = (1-\alpha)/(1-\alpha+\beta)$, it has a second pure-strategy equilibrium in which neither state attacks.*[14] *A state's ex ante expected payoff in the more peaceful equilibrium is* $1 - \epsilon(1+\beta)$.

If the risk of accident or an aggressive type is high enough, peace become impossible as the players are motivated to avoid getting the "sucker" payoff β. $\underline{\epsilon}$ is exactly the probability of attack that makes a player in the Stag Hunt of Figure 1 indifferent between attack and don't attack. The game with no signals is thus effectively equivalent to the simultaneous move Stag Hunt with an ϵ probability of crazy types or accidents.

If the signals are perfect – no Type I (false positive) or Type II (false negative) errors – then we can show

[13] Proofs are in the Appendix.

[14] Here and throughout "state" will generically refer to the non-crazy types that have a choice to make.

Proposition 2. *With perfect signals, the game always has a pure-strategy equilibrium in which both states attack for sure, regardless of the signal they see. If* $\epsilon < \bar{\epsilon} = (2-\alpha)/(2-\alpha+\beta)$ *it has a second pure-strategy equilibrium in which neither state attacks when it sees the signal "no attack" and attacks otherwise. Ex ante expected payoffs in this more peaceful equilibrium are* $1-\epsilon(1+\beta/2)$*, which is* $\epsilon\beta/2$ *greater than in the no-signals case.*

Notice that $\underline{\epsilon} < \bar{\epsilon}$, so that peace can be supported for a higher ex ante probabilities of accident or crazy types when signals are perfectly informative. The reason is that in the event that a state is the second mover, with perfect signals it will know if the other attacked because it was an aggressive type, and so can ensure getting 0 instead of $-\beta$. Perfect warning protects somewhat against the risk of getting the worst, "second strike" payoff. As we will see, with imperfect warning this benefit comes along with the downside of war risk due to false positive of an attack.

A second interesting point is that with perfect signals, we can support an equilibrium in which the states don't attack even when $\alpha > 1$, which is to say that they prefer first-strike to peace! Notice that $\bar{\epsilon} > 0$ for $\alpha \in (1,2)$, so with $\epsilon < \bar{\epsilon}$, (NA, NA) can still be an equilibrium path for α in this range.[15] This is surprising, since intuitively one might have thought that with $\alpha > 1$ attacking must dominate. One might reason that with $\alpha > 1$, the second mover would always want to attack; anticipating this the first mover would want to attack to avoid getting $-\beta$. If you would want to attack in either position, you might reason that you should attack regardless.

But this is not right. Here, a player who sees the signal "no attack" *can't be sure they are the second mover* and thus able to get away with a first strike. If they are not, then attacking will cause the other to attack, which is a worse outcome than neither attacking, even when $\alpha > 1$.

There is an important point here that generalizes to the case of imperfect signals: Good warning technologies have a *deterrence effect*, by lowering the odds that a player can get away with a first strike.

[15] By sharp contrast, in the game shown in Figure 1, $\alpha > 1$ makes it a Prisoners' Dilemma, with *Attack* a dominant strategy for both players. Similarly, the game here with no signals has mutual attack as the unique equilibrium when $\alpha > 1$.

A related interesting result from the perfect-signals case is that perfect signals do not get rid of the "attack whatever signal you get" equilibrium. In this equilibrium, the first mover gets the "no attack" signal and therefore knows for sure the other hasn't started its attack. So why not choose not attack, to convince the other there is no threat? The problem is that on seeing "not attack" the second mover would conclude that *it* is the first mover, which undermines the ability to convey "no attack" even though the signal is perfect (no noise).[16]

The dilemma arises because, realistically, the players do not know if the other has already made its choice or not. Seeing "no attack" tells you that there is no attack happening now (perfect signal) but does not tell you if the other side is about to launch an attack. If you think that they think that you are going to attack, then your not attacking will not forestall their attack, because they will just think they are the first mover and have to attack because you are about to.

The effect is similar to simultaneous choice as in Figure 1. If instead it was common knowledge that state 1 chooses before state 2 then there is no equilibrium in which both attack (if $\epsilon < \underline{\epsilon}$). In that game structure state 2 knows it is not being attacked and knows it is safe if it does not attack, so its sequentially rational choice is to not attack (to get 1 instead of α). State 1 can anticipate this so it chooses not to attack to begin with.[17] Here by contrast, if mutual attacks are expected, then not attacking does not provide a clear signal to the other that they are safe to not attack, because they will just infer that you haven't started your attack yet.

5 Noisy signals

Recall that in the imperfect signals case, each state observes a number z_k, where z_k is drawn from the cumulative distribution F_0 if there is no attack in progress, and from F_1 if there is an attack in progress. Given signal z_k, the likelihood ratio of an attack is $L(z_k) = f_1(z_k)/f_0(z_k)$, which is assumed to be increasing in z_k, so that higher signals are more likely when there is actually

[16] Indeed, somewhat remarkably $(Attack, Attack)$ is an equilibrium for any ϵ including $\epsilon = 0$.

[17] Although not shown in this draft version, this results holds in the infinite horizon sequential version of the game in which states alternate in choosing whether to attack. There, deviating to not attack puts off war with payoffs of (0, 0) for a period, which makes the off-path best reply to not attack also not attack.

an attack happening. The greater the variance of the distributions F_1 and F_0, the flatter $L(z_k)$, meaning basically that signals are less informative.

Substantively, higher z corresponds to warnings from a state's warning systems that are considered more concerning and more indicative of either preparations for nuclear attack, or an actual attack having begun.

Because higher signals always cause higher beliefs that an attack is underway or about to begin, we can conjecture that in any equilibrium, if a state attacks on signal z_k it also attacks for any signal $z'_k > z_k$. So we will consider possible equilibria that have the form, "attack if and only if the signal you get is greater than $\hat{z}_k$," where $\hat{z}_k$ is a cutpoint, or threshold. The lower a state's $\hat{z}_k$, the *more* of a "hair trigger" the state is on, since this means that it will attack for lower signals.

To begin with some intuition, suppose that you have received a somewhat concerning signal. If you initiate an attack, you will end up with a payoff of either α or zero. You get α if the other side was the first mover and didn't attack (so you attacked on a "false positive"), or if you are the first mover and they do not launch after seeing whatever signal your attack actions generated for them (they get a false negative). By contrast, if you choose not to attack, you end up with a payoff of either 1 or $-\beta$. You get 1, the peace payoff, either if the other was first mover and was not actually attacking, or if you are the first mover and they do not attack on the signal that your not attacking generates for them.

If the other state's threshold is $\hat{z}$, then given that you see signal z, attacking is better than not attacking when

$$\begin{aligned} u(\text{Attack}|z,\hat{z}) &> u(\text{Don't attack}|z,\hat{z}) \\ \alpha G(z,\hat{z}) + 0 * (1 - G(z,\hat{z}) &> 1 * S(z,\hat{z}) - \beta(1 - S(z,\hat{z})) \\ \alpha G(z,\hat{z}) &> (1+\beta)S(z,\hat{z}) - \beta. \end{aligned} \tag{1}$$

Here, $G(z,\hat{z})$ is the probability that a state that chooses to attack *Gets away with* a first strike when it sees signal z and the other state's threshold is $\hat{z}$. Likewise, $S(z,\hat{z})$ is the probability that the state would be *Safe* (not attacked) if it chooses to not attack, given the signal z and the

other state's threshold at $\hat{z}$.

To anticipate, a Bayesian equilibrium will be defined by any $z = \hat{z}$ that makes (1) an equality. The difficulty is in specifying the G and S functions, which are somewhat involved. The reason is that they depend on the states' posterior (post-signal) beliefs about which of three "states of the world" obtain. To avoid the confusion with "state" as player, I will call these *conditions* $\omega \in \{1, 2, 3\}$.

In the first condition, $\omega = 1$, the player is the first mover and thus no attack is in progress. This condition has prior (pre-signal) probability of one half, since each is equally likely to be first mover. In the second condition, the player is second mover and there is no attack in progress. This condition has prior probability $(1 - a_i)(1 - \epsilon)/2$, where a_i here can be the probability that the first mover initiated an attack. In the third condition, the player is the second mover and there *is* an attack in progress. This condition has prior probability $((1 - \epsilon)a_i + \epsilon)/2$.

On observing signal z, the state can update about which condition they are in. Intuitively, if you see a high signal, your posterior belief that you are the second mover and an attack is underway should increase (condition 3) relative to no attack (conditions 1 and 2). A low signal should increase your belief that no attack is in progress (conditions 1 and 2 relative to 3).

Table 1 shows the prior beliefs, likelihoods, and posterior probabilities for the general case of signals with densities f_0 and f_1, and a_i the expected action of a non-crazy other state (which we can allow to be a probability in between 0 and 1 here). The last row before the definitions rewrites the posterior probabilities of the three conditions in terms of the likelihood ratio of the signal, $L(z)$.

The expressions in Table 1 make it clear that a player's pre- and post-signal beliefs about what is happening should rationally depend not only on the signal by itself, but also on their beliefs about the other player's strategy (a_i in the Table). For example, the less one expects the other to be inclined to attack (lower a_i's), the smaller the impact of any given concerning signal (higher z) on one's belief that an attack is in progress. Though contrary to a strong tendency in the literature on surprise attacks – which often takes people to task for putting weight on prior beliefs – this is just a consequences of Bayes' rule. The signal is not the only thing that one should rationally

Table 1: Priors and posteriors on observing signal z

	Condition (state of the world)		
	$\omega = 1$ (1st mover)	$\omega = 2$ (2nd mover, no attack)	$\omega = 3$ (2nd mover, attack)
Prior belief	$1/2$	$(1-\epsilon)(1-a_i)/2$	$((1-\epsilon)a_i + \epsilon)/2$
Likelihood of z	$f_0(z)$	$f_0(z)$	$f_1(z)$
Posterior belief (by Bayes' rule)	$\frac{f_0(z)}{2D}$	$\frac{f_0(z)(1-\epsilon)(1-a_i)}{2D}$	$\frac{f_1(z)[(1-\epsilon)a_i+\epsilon]}{2D}$
Posterior belief (rewritten)	$p_1(z) = \frac{1}{E}$	$p_2(z) = \frac{(1-\epsilon)(1-a_i)}{E}$	$p_3(z) = L(z)\frac{(1-\epsilon)a_i+\epsilon}{E}$
Definitions:	$D = Pr(z) = f_0(z)/2 + f_0(z)(1-\epsilon)(1-a_i)/2 + f_1(z)[(1-\epsilon)a_i + \epsilon]/2$		
	$E = 1 + (1-\epsilon)(1-a_i) + L(z)[(1-\epsilon)a_i + \epsilon]$		
	$L(z) = f_1(z)/f_0(z)$		
	$a_i = \Pr(\text{other state attacks if 1st mover}) = 1 - F_0(\hat{z})$ if $\hat{z}$ is threshold		

consider when making a judgement about whether an attack is happening.

With threshold strategies, a state attacks with probability $1 - F_0(\hat{z})$ if it is not under attack (conditions 1 and 2), and with probability $1 - F_1(\hat{z})$ if an attack is in progress. So, at last, we can specify G and S, as follows:[18]

$$G(z, \hat{z}) = p_1(z)(1-\epsilon)F_1(\hat{z}) + p_2(z) = \Pr(\text{Get away with first strike}|z, \hat{z})$$

$$S(z, \hat{z}) = p_1(z)(1-\epsilon)F_0(\hat{z}) + p_2(z) = \Pr(\text{Safe if you don't attack}|z, \hat{z}).$$

A characterization of equilibria follows.

Proposition 3. *The strategy "attack for any signal" is always an equilibrium. The strategy "do not attack for any signal" cannot be an equilibrium for any warning technology that has* $\lim_{z\to\infty} L(z) >$

[18] In these expressions, $p_1(z)$ and $p_2(z)$ use $a_i = 1 - F_0(z)$ in the E denominators shown in Table 1.

$2(1-\epsilon)(1-\alpha)/\beta\epsilon - 1$ *(which includes Normal distributions, which have* $\lim_{z\to\infty} L(z) = \infty$*). If* $1+\beta > \alpha$*, then "attack if* $z \geq \hat{z}$ *and do not attack if* $z < \hat{z}$*" is an equilibrium for any finite* $\hat{z}$ *that satisfies*

$$\alpha G(\hat{z}, \hat{z}) = (1+\beta)S(\hat{z}, \hat{z}) - \beta. \tag{2}$$

With some algebra, this equilibrium condition can be rewritten more usefully if less simply in terms of a state's *best-reply threshold* $z_{br}(\hat{z})$ when the other has threshold $\hat{z}$. Let $z_{br}(\hat{z})$ be the cut point such that a state would want to attack on seeing $z > z_{br}$ and not attack for $z < z_{br}$ when it thinks the other state's "trigger" is set to $\hat{z}$. We have

Proposition 4. *For general warning system technologies* F_0 *and* F_1*, a state's best-reply threshold given* $\hat{z}$ *is defined by any* z_{br} *that satisfies*

$$L(z_{br}) = R(\hat{z}) \equiv \frac{(2-\alpha)F_0(\hat{z}) - \alpha F_1(\hat{z})}{\beta\left(\frac{1}{1-\epsilon} - F_0(\hat{z})\right)} - 1. \tag{3}$$

For $F_a(z) \sim N(a, \sigma^2), a \in \{0, 1\}$*, this solves to*

$$z_{br} = \frac{1}{2} + \sigma^2 \log R(\hat{z}). \tag{4}$$

The condition for an equilibrium with positive probability of peace is then $\hat{z} = z_{br}(\hat{z})$*. If an equilibrium of this form exists, then the largest and thus most peaceful one is "stable" in the sense that* $z'_{br}(\hat{z}) < 1$*.*[19]

Figure 2 shows examples of the best-reply function $1/2 + \sigma^2 \log R(\hat{z})$. Equilibria are where this curve crosses the 45 degree line. It is apparent that it can cross either twice or not at all (generically).[20]

Zero intersections corresponds to a situation in which the unique equilibrium has both attacking regardless of the signal. The case shown ($\alpha = \beta = .68$, $\epsilon = .3$, $\sigma = .7$) illustrates a pathology

[19] In the more general case, $L'(\hat{z}) > R'(\hat{z})$, meaning that L cuts R from below.

[20] In the proof of Proposition 4 I show that for Normal F_a, $R(z)$ is unimodal, increasing to a unique maximum and then decreasing. This guarantees either zero or two intersections of $L(z)$ and $R(z)$ in the Normal case (generically), and, somewhat more generally, for any F_a whose likelihood ratio function is weakly convex.

of sorts. With these parameters, if the parties could commit ex ante to have *no* warning system at all, or to completely ignore their signals, they could support the strongly mutually preferable equilibrium in which neither non-crazy types attacks, despite the .3 chance the other is the aggressive type.

One could say that in this case a *reciprocal fear of surprise attack* drives a bad outcome. Here, for any given warning system sensitivity (i.e., threshold), the chance that the other player will see a signal that makes them want to "shoot" is high and dangerous enough that you want to set an even more "hair trigger" threshold yourself. In consequence both just "draw" straight away, for any signal.

Two intersections corresponds to one stable and one unstable equilibrium, at the higher and lower intersections, respectively. "Stable" means that if a player expected that the other was using a nearby threshold, say $z^* = z_{br}(z^*)$ plus or minus a small amount, then the player's best reply would be to move in the direction of z^* (the equilibrium). At the unstable equilibrium, small departures from z^* make for best replies that move away, towards either the higher (more peaceful) equilibrium or downwards towards the $(Attack, Attack)$ equilibrium.

Next, notice that the best-reply function is increasing to a maximum point, and then decreases towards a limiting value. Where it is increasing, the states' choices of warning system sensitivity are *strategic complements*, as Schelling conjectured. This means that in this range, the more relaxed (or hair trigger) about signals the other state is, the more relaxed (or hair trigger) you want to be.

For intuition, note that a state's choice of warning system sensitivity involves a trade off between the risk of a false positive – the probability that you attack when not under attack, $1 - F_0(\hat{z})$ – and the risk of a false negative – the probability that you don't attack when under attack, $F_1(\hat{z})$. Making your warning system more "hair trigger" (lowering $\hat{z}$) increases the risk of war from a false positive, while reducing the risk of war from a false negative.

In the range of $\hat{z}$ where $z_{br}(\hat{z})$ is increasing, if you make your warning system less hair trigger, this makes the other side less worried that they will be attacked due to you getting a false positive signal. This makes them more relaxed about signals that previously would have been more

concerning.

Interestingly, however, if you think the other side has a sufficiently high $\hat{z}$ (is very relaxed), then increases in their $\hat{z}$ (more relaxed) make you want to make your warning system *more sensitive* (strategic substitutes). If the other side is very relaxed, there is almost zero chance of war with a non-crazy type due to their getting a false positive, and if you make your warning system a little more sensitive this will have almost no impact on the other side (because the chance of a false positive is already close to zero). However, making your warning system a little more sensitive *does* then protect you a little better if the other side happens to be a crazy type, or launches accidentally, by reducing the chance of a false negative.

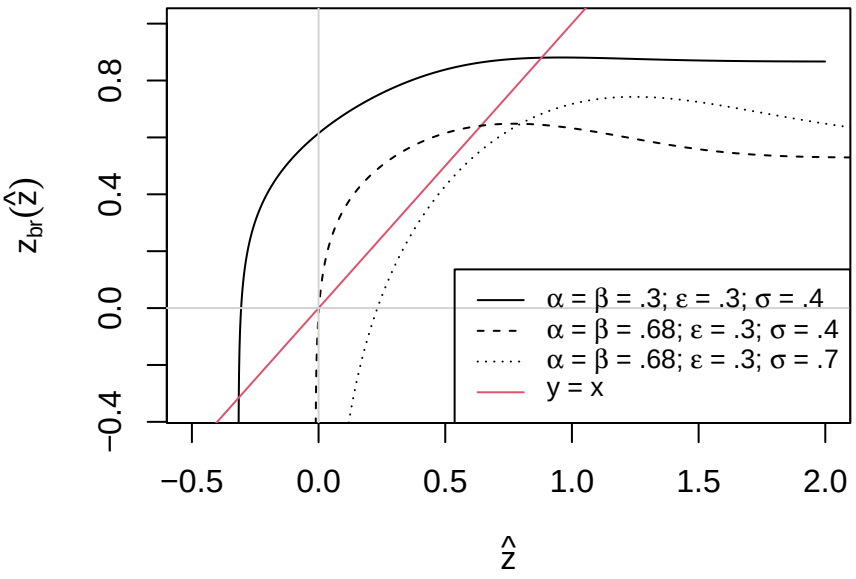
In brief, if there was zero chance of accidents or crazy types, states would ideally like to ignore warning entirely, confident that neither has any incentive to attack the other. But positive chance of accident or an aggressive type means that there is some value in warning, and the more so the greater the difference between first and second-strike, or, say, accident risk in a crisis or war. In turn, however, noisy warning means that now attack may be triggered even absent an accident or aggressive type, so that states want to make their thresholds somewhat more sensitive (hair trigger) to protect against this possibility. Doing so with one's own system has the "negative externality" of making the other more nervous about your false positives, leading to Prisoners' Dilemma-type of problem, as discussed next.

6 Escalation risk

When $\alpha < 1$, either player attacking is ex ante inefficient – a bad thing – for the non-crazy types.[21] As a shorthand I will use *peace* to refer to the outcome where neither attacks, and *war* for outcomes where at least one player attacks. War is more likely when z^* is lower, which means that warning system sensitivities are higher. Peace is more likely the more that the players discount higher

[21] When $\alpha \in [1, 2]$ any attack is inefficient in the weaker sense that peace (no attack) maximizes the sum of player payoffs.

Figure 2: Warning system sensitivity, best reply functions and equilibrium



signals, meaning higher z^*.[22]

Proposition 5 describes how the chance of attack varies with parameters α, β, ϵ, and σ in a stable equilibrium where peace may occur. It seems most reasonable to focus on peaceful equilibria given that peace would be the focal or default expectation based on the players starting the game – or this moment of it – at peace.

Proposition 5. *At any stable equilibrium in which peace has positive probability, increasing the first-strike advantage α, the second-strike disadvantage β, or the chance of an accident or crazy type ϵ, increases the likelihood of war. For large enough $\beta > 0$ or for large enough $\epsilon \in (0, 1)$, there is no peaceful equilibrium.*

Regarding how noisy the warning systems are, for the case of $F_a \sim N(a, \sigma^2)$, in the maximal peaceful equilibrium, increasing σ from zero initially increases the risk of war, and then for large enough σ, further increases decrease the risk of war. It is possible for the peaceful equilibrium to exist for low and high σ, while only the "attack for any signal" equilibrium exists for an intermediate range of warning system noisiness σ.

For α, β, and ϵ, these comparative static results follow from the fact that increasing any of them shifts $R(z)$ down, which then implies lower equilibrium $\hat{z}$ (see Figure 2). Greater first-strike advantage or second-strike disadvantage makes tragic escalation more likely by leading the players to attack for less diagnostic signals, because getting away with a first strike is better, or suffering a second strike is worse. Increasing the prior belief that the other player is an aggressive type – or that an accident might occur, as during a crisis – means that the same signal becomes more concerning, because one's prior was already higher. This is Bayes' rule in action.

Although these results are intuitive, note that in the simple Stag Hunt model of Figure 1 (and other generic normal-form coordination games) a small change in payoffs has no effect at all on a player's equilibrium strategies in any pure-strategy equilibrium. An unsatisfactory aspect of Nash

[22]The ex ante (pre-signal) chance that neither player attacks in equilibrium is $(1-\epsilon)^2 F_0(z^*)^2$, and the chance that at least one attacks is one minus this. The probability of war conditional on both types being non-crazy is similarly $1 - F_0(z^*)^2$.

equilibrium analysis in coordination games is that we have strong intuitions about how variation in payoffs will influence players' propensities to take different actions, perhaps by affecting their expectations about which equilibrium to expect. While the analysis here stays within the "focus on one equilibrium" paradigm, by enriching the coordination problem with noisy signals of intended actions, we arrive at an intuitively natural, smooth and not-flat relationship between payoffs and propensity to choose attack or not attack.

Another striking contrast is that in the simple Stag Hunt game with ϵ chance of accident, neither attacking is an equilibrium when $\epsilon < \underline{\epsilon} = (1-\alpha)/(1-\alpha+\beta)$, whereas here it is entirely possible for "both attack" to be the only equilibrium even when this condition holds. In other words, noisy warning introduces a Prisoners' Dilemma aspect to the underlying coordination problem. For $\epsilon < 1/(1+\beta)$, both states could be better off if they could commit to have no warning systems or to ignore all signals, leading to an expected payoff of $1-\epsilon-\epsilon\beta > 0$. This is better than the zero payoff they get in "both attack." But each has an incentive to use warning to protect itself against accident or crazy-type. This creates a risk of false positives that then feeds back on the other side's incentive to choose a lower (more sensitive) threshold.

More surprising comparative statics results relate to σ, which measures how imprecise the warning systems are. Figure 3 illustrates how the equilibrium probability of war varies with signal precision for different levels of first-strike advantage/second-strike disadvantage, in the Normal F_a case.

In this example $\epsilon = .1$, so if neither non-crazy type were to attack for any signal, then the probability of war would be $1-(1-.1)^2 = .19$. Notice that for any degree of first-strike advantage, the probability of war approaches .19 for very precise and very noisy signals, as in the analysis of the no-signal and perfect-signal cases in section 3. In between, we see that the chance of war is upside-down U-shaped. However,

a. for small first-strike advantage (e.g., $\alpha = \beta = .1$ here), the chance of war is basically .19 for *any* signal precision, and

b. for large-enough first-strike advantage, there is a range of σ where the maximally peaceful equilibrium no longer exists, and the unique equilibrium has both sides attacking for sure (thus, probability of war is one).[23] As first-strike advantage increases further, the range of σ with $P(war) = 1$ gets larger.

The case of the so-so warning system in (b) is clearly pathological. The players would be *much* better off if they could commit to having no warning system at all (same as very large σ). With a so-so warning technology, all thresholds $\hat{z}$ imply either a false positive or a false negative rate high enough that it is better to just attack. Mutual commitment to a high threshold is not feasible because the best reply to the other being very relaxed is to edge up one's own rate of detecting attacks correctly, which unfortunately also increases your false positive rate, which ramifies back on the other side's incentive to get more hair trigger.

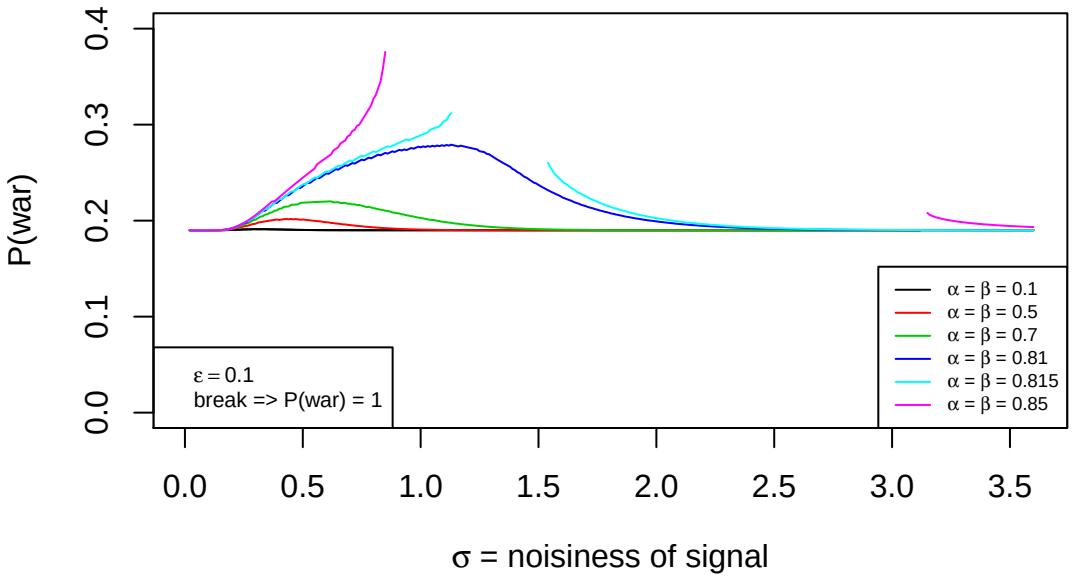
Fortunately, this pathology requires what appear to be large first-strike advantages, and non-trivial risk of crazy types or accidents. If any nuclear war is judged to be extremely bad by state leaderships – e.g., MAD is not in doubt – this implies that α and β are small, in which case equilibrium has both sides setting very relaxed (high) thresholds that send the false positive rate to zero and the false negative rate to one.

7 Equilibrium versus the first best

The states have a strong common interest in avoiding war due to false warning. But as we have seen, they can also have a unilateral incentive to edge down their threshold for attack to reduce the false negative rate, which unfortunately has the effect of increasing the other side's danger from a false positive (the negative externality). This Prisoners' Dilemma-logic inherent in noisy warning systems makes for higher risk of inadvertent escalation than if the states could somehow commit to less attack sensitivity to warning.

[23] All of these examples satisfy $\epsilon < \underline{\epsilon}$, so the failure of the peaceful equilibrium is due to the introduction of a "so-so" signaling technology.

Figure 3: Signal precision and the probability of attack



To assess how much added risk there is, and under what circumstances it is large, we need to compare the Bayesian Nash equilibrium threshold at the stable equilibrium, $\hat{z}$, to the first-best choice of threshold, z^{fb}. The latter is the threshold that maximizes the states' ex ante expected payoffs if both could commit to implement it in their attack decisions (something I expect is impossible). Note that as a general matter the first-best does *not* entail commitments to not attack for any signal whatsoever – the states have a common interest in protection against accidental and crazy-type attacks.

Figures 4 and 5 display examples of how the states' equilibrium payoffs vary and compare with the first best as we vary the the noisiness of the warning system and the size of the first-strike advantage by itself ($\alpha = \beta$ here).[24]

For σ, we see that when first-strike advantage is large enough, "so-so" warning system accuracy can completely eliminate any peaceful equilibrium – note that payoffs drop to zero where the green line is broken. When first-strike advantage is small, by contrast, there is hardly any risk of inadvertent escalation due to false alarms between non-crazy types, regardless of how good or bad the warning system is.

For first-strike advantage α (which here I am equating with second-strike disadvantage on the down side), the departure from the straight line represents the additional loss beyond the loss from risk of accident or crazy type. Here there is an interesting non-monotonicity in σ: For very accurate or very inaccurate warning systems, there is not much risk of inadvertent escalation even for relatively large first-strike advantages. Again, the negative-externality problem arises for so-so warning systems, getting worse and worse as first-strike advantage increases.

[more here . . .]

[24] z^{fb} does not have a convenient closed form solution. It is the z that maximizes a state's ex ante expected payoff, which is $u(z) = -\epsilon\beta(F_0 + F_1)/2 + (1-\epsilon)[F_0^2 - F_0(1-F_0)\beta + (\alpha-\beta)F_1(1-F_0))]$. ($F_0$ and F_1 are functions of z here.) I believe it is possible to show that $z^{fb} > \hat{z}$ when a peaceful equilibrium exists, but am still working on it.

Figure 4: Warning system accuracy and expected payoffs

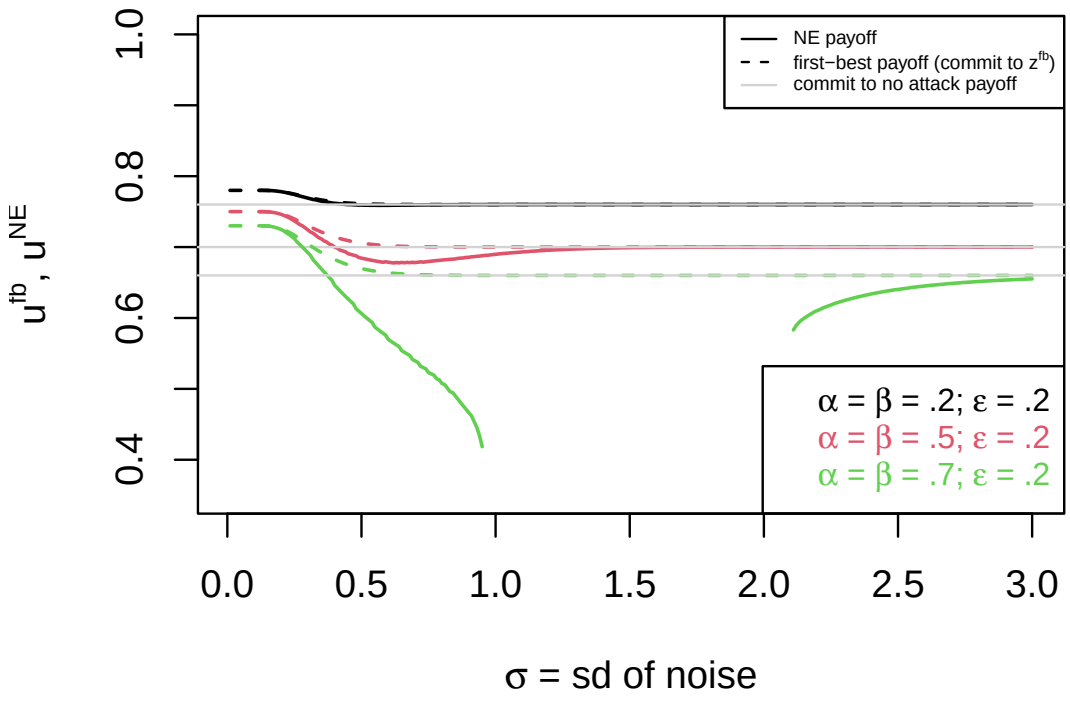
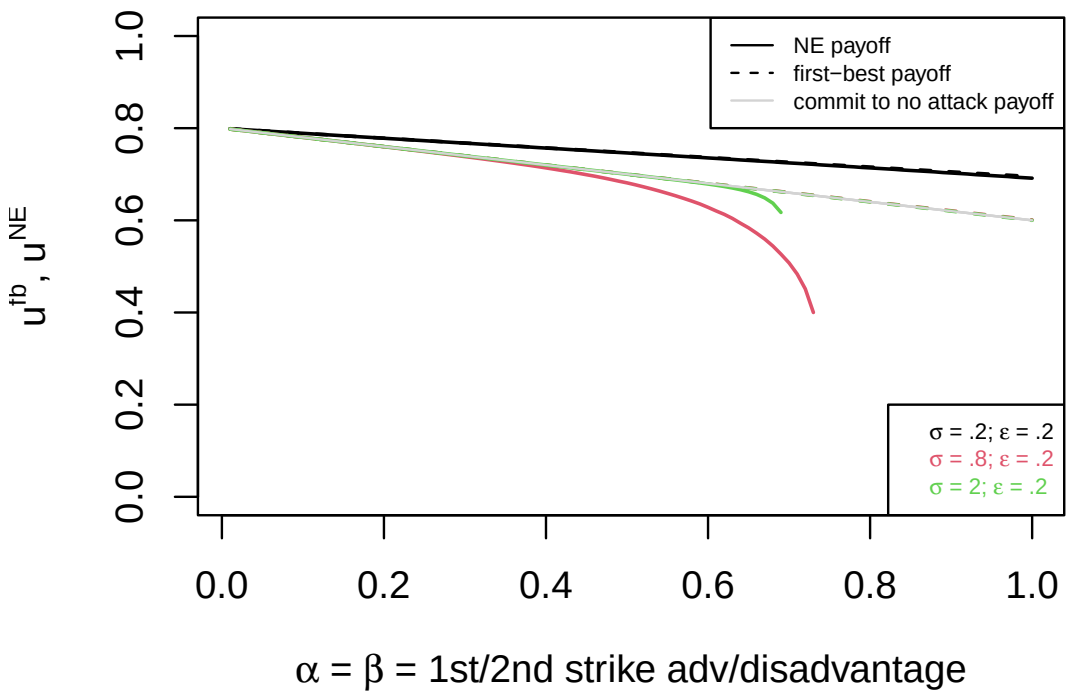


Figure 5: 1st/2nd strike adv/disadvantage and expected payoffs



8 Surprise attacks and intelligence failures

An interesting and perhaps surprising implication of the analysis above is that in this signal-augmented Stag Hunt, *players' payoffs strongly influence how they interpret new evidence about the other's behavior*. This is not an irrational mechanism like wishful thinking or motivated bias, which are linked to failures to use Bayes' rule properly. Instead the connection arises because in this strategic situation, how a player updates from a signal depends on their prior belief about the likelihood of attack, which in turn depends on their equilibrium expectations, which in turn incorporate beliefs about how the other player will evaluate signals.

For example, if first-strike advantages are small, then in the maximally peaceful equilibrium the players will *endogenously* strongly discount even highly diagnostic signals of attack. Suppose that the chance the other is a crazy type is 10%, signal standard deviation $\sigma = .5$, and the first-strike advantage and second-strike disadvantage is .1 ($\alpha = \beta = .1$). Then in the most peaceful equilibrium, $z^* = 1.78$, or 3.56 on a standard Normal distribution. This implies that the probability of a false positive – i.e., that a player gets a signal that leads it to attack when it is not actually under attack itself – is just .000019, basically zero. But the probability of a false negative – the probability that a state gets a signal that leads it to *not* attack when it *is* under attack – is .94! So the chance that a state disregards a correct signal that it is under attack is approximately $.1 * .94 = .094$, or close to 10%. The minimum likelihood ratio needed for a signal to lead a (non-crazy) state to attack is $L(z^*) = 167$, implying that one should require extremely compelling evidence even though there is a non-trivial chance of being attacked.

How can this be? In this case, before seeing the signal a state's prior belief that it is the second mover and is under attack is approximately $.1/2 = .05$ (since the other player is almost certain to not attack if it is non-crazy or doesn't have an accident). And because the first-strike advantage is small relative to remaining at peace, a state prefers to Not Attack unless it thinks there is at least a 90% chance that it is under attack or will be attacked. To get from a prior of .05 to a posterior of .90, one needs a signal with a likelihood ratio of about 167.

The situation is very different if payoffs are such that players want to attack if they think there is at least, say, a 50% chance that they are under attack or will be attacked. Then a less diagnostic signal will make one want to attack, and in turn the player's must worry much more about the other getting a false positive, which feeds back on their inclination to credit less clear signals.[25]

After Pearl Harbor and after 9/11, scholars and high-level commissions appropriately undertook extensive investigations into the reasons for the intelligence failures (e.g., Wohlstetter, 1962; Kean, 2011). In both cases analysts found many warnings that, ex post, seemed like they should have been relatively clear indications that an attack was coming (indications treated as false negatives). A major issue in both cases was said to be a bureaucratic failure to "put the pieces together" due to stove-piping and other organizational glitches or pathologies. This is not exactly the same as having a very high threshold for concluding that attack is likely enough to merit action. But arguably the propensity of parts of the intelligence bureaucracy to flag concerning signals for superiors is a functional equivalent of $\hat{z}$ in the model. That is, what an intelligence analyst considers worth flagging and asking for more attention from higher ups is function of the system's overall "setting" for warning. Similarly, the higher-ups propensity to search for ambiguous warning from their bureaucracy will also be a function of the implicit $\hat{z}$, as are their decisions about what warning reports to act on.

If so, then the analysis here might help us to better understand the difficult tradeoff that lies at the heart of the problem of warning against surprise attack. If attack is thought not to be highly likely – definitionally the case for "surprise attacks" – then it makes good strategic sense to set one's warning system to strongly discount signals of possible attack. When preemptive actions might themselves trigger a bad outcome, it can make sense to accept a high risk of false negatives to avoid the more immediate danger and risk from false positives. (Granted this is more relevant for Pearl Harbor-type cases than 9/11 type cases.)

[25]Observation: From $R(\hat{z})$, it is immediate that for any σ and $\epsilon > 0$, $\lim_{\alpha,\beta\to 0} z^* = \infty$, so that the probability of a false positive goes to zero and a false negative goes to 1 as first-strike advantages get small.

As a result, when a surprise attack occurs, it is also highly likely that there will have been relatively clear warnings signs that were not acted on.

Ask yourself, do we want to reduce the risk of a surprise nuclear attack by making our launch threshold for warning much lower? This would reduce the risk of a false negative, but at the same increase the risk of triggering war by a false positive.

9 Asymmetric first- and second-strike capabilities

Recall the Hawaii missile alert and its political context: the Trump administration was apparently seriously considering a first-strike effort to eliminate North Korea's small-but-growing nuclear capabilities and, almost surely, the regime. It is very likely that North Korea's leadership understood and greatly feared the implications of the administration believing that it might be able to execute a successful, disarming first strike. Here, the U.S.'s expected payoff for a first-strike attempt was plausibly much larger than North Korea's, and North Korea's second-strike ability to inflict damage on the U.S. much smaller that the U.S.'s second-strike ability to damage North Korea and its regime. Though much less extreme, the Cuban Missile Crisis also played out in a situation of asymmetric capability to inflict nuclear damage.

Or consider the police-encounter problem, where surely police are typically far more concerned about "crazy types" than about non-crazy types shooting to prevent being shot. To a first approximation, a better model of this context may be decision-theoretic, with crazy types shooting and non-crazy types not shooting with given probabilities ϵ and $1-\epsilon$ (thus assuming that non-crazy types never shoot, but instead try to surrender or run away).

This section briefly describes results for the model with asymmetric payoffs, (α_1, β_1) for state 1 and (α_2, β_2) for state 2.

[results/figures to come. Main point: Asymmetry can *exacerbate* inadvertent escalation risk and lower the expected payoffs of the ostensibly advantaged side. For example, increasing α_1 holding all other payoffs fixed can *decrease* state 1's expected payoff by driving down both states' attack thresholds. Greater α_1 makes state 1 willing to attack for less clear warning, which in turn

makes state 2 more concerned about false negatives, which ramifies back on state 1, and so on.

These results speak to a long-standing and newly invigorated U.S. policy debate on the value of pursuing *counterforce* nuclear capability, meaning capabilities that reduce another nuclear state's ability to impose damage on the U.S. by preemptive attack on their launch platforms and decision systems. Advocates of pursuing counterforce first-strike potential make several arguments:

1. That it is the responsibility of the military to defend and protect U.S. citizens, so that "damage limitation" by preemptive counterforce is a requirement.

2. Visibly strong preemptive counterforce capabilities give the U.S. an advantage in resolve in international crises and wars with other nuclear-weapons states, and so would also contribute to ex ante deterrence.

3. In particular, the U.S. needs to pursue preemptive counterforce to reassure allies and partners about the credibility of U.S. extended nuclear deterrence.

4. That it is immoral to attack non-military targets (McNamara 1962 "no cities" speech). This consideration does not have to entail going for potential first-strike capability, but at least could entail limits on second-strike damage.

With respect to the model, strengthening preemptive counterforce is equivalent to increasing α_1 without changing β_1 (or maybe even decreasing it). But this can actually *reduce* state 1's ex ante expected payoff by increasing crisis instability – the probability of inadvertent escalation goes up enough to overwhelm the effect of increasing one's "get away with first-strike" payoff. Further, deterrence is not necessarily strengthened, or effects would be minimal relative to issues at stake in a militarized dispute, because crises are getting more dangerous for *both* states.

...

]

10 Conclusion

This paper examined a coordination problem in which the players think there is a (possibly very small) chance the other side will attack no matter what, and in which they observe a noisy signal of whether the other player has already begun an attack. The non-aggressive types in the game would prefer to have no attacks (peace) to any attack (war), but if under attack it is best not to delay. I labelled and interpreted it as a model of the "reciprocal fear of surprise attack" problem that Schelling (1960) discussed for the case of nuclear war. But the same underlying strategic problem appears in a range of other important settings, such as police encounters and the coup/purge dynamics of elite politics in weakly institutionalized states. Indeed, in those settings the first-strike advantages are probably on average much stronger than in the nuclear case, at least between states for which mutual assured destruction obtains.

One main finding is that how the players interpret and act on signals is heavily dependent on their equilibrium expectations and also on the payoffs for different outcomes. When the first-strike advantage is small, players tend to dismiss all but extremely compelling signals, so that the probability of a false positive is very low while the chance of false negative if there is an attack (i.e., dismissing a valid signal) is close to one. Tragic escalation is very unlikely in the most peaceful equilibrium. By contrast, with larger first-strike advantages, players endogenously adopt more "hair triggers," which increases the odds that one will attack in the mistaken belief that they are under attack. Because hair triggers prove to be (mainly) strategic complements, when the signal technology is neither highly accurate nor highly inaccurate, the players go to a level of sensitivity that is suboptimally too high. They would both better off, and tragic escalation would be less likely, if they could commit to be somewhat more relaxed about signals of possible attack. Indeed, in extreme cases of very strong first strike advantage, a "so-so" signaling technology can actually eliminate all but the bad equilibrium in which both sides immediately attack no matter what signal they get.

Some natural extensions would involve articulating the model for specific settings that differ

from the symmetric nuclear war scenario. For example, for police encounters it seems reasonable to consider asymmetries. If a suspect in fact has no gun, damage limitation is not an option, so the problem is arguably one-sided in a way that might undercut or make for a different kind of reciprocal fears dynamic.

Even in the case of nuclear war, the situation of the US and North Korea differs consequentially from the US vs. USSR case. Kim Jong Un may reasonably suspect that the US leaders think they could undertake a disarming first strike, whereas it is clear that the US can strike back. In other words, assured destruction in not yet mutual. How would this asymmetry change things in the model?

It is wildly implausible to imagine police and suspects making precise Bayesian calculations in their split-second, high intensity interactions. Even in the nuclear and coup/purge cases, time can be short and pressures extremely high. So, as is normally the case, the model should not be taken as a literal depiction or account. It is instead a way of better understanding a type of strategic situation by clarifying its structure and what this implies for rational decision-makers. Ex ante, it is not completely obvious just what these three situations have in common and just how they differ in terms of implications for rational behavior and the welfare of the players. It *is* plausible that in the real world people's choices in these situation reflect both prior beliefs (which can in turn reflect various biases) and updating from new information (see for good examples of this Aveni, 2008). The results here may be helpful by showing how priors are endogenous to equilibrium expectations, and decision rules for responding to new information are strongly influenced by payoffs (e.g., size of first-strike advantages). The model's implications concerning these factors are arguably quite reasonable, and do a good job of matching different typical patterns of behavior and inference across the different settings of coups, police encounters, and nuclear stability.

11 Appendix

Proof of Proposition 1. With no signals, a complete strategy is just a choice of attack or not attack. If the other state is expected to attack, attack yields 0 (regardless of whether one is first or second mover) and not attack yields $-\beta$, so both attacking is an equilibrium. If the other state is expected to not attack, then attack yields $(1-\epsilon)\alpha + \epsilon 0$ and not attack gives $1-\epsilon+\epsilon(-\beta)$. The comparison leads to the condition in the Proposition on ϵ.

Proof of Proposition 2. With perfect signals, if a state sees the signal "no attack" and expects that the other non-crazy type chooses not to attack if it sees "no attack," then its posterior beliefs are

$$Pr(\omega = 1) = \frac{1}{2-\epsilon}, Pr(\omega = 2) = \frac{1-\epsilon}{2-\epsilon}, Pr(\omega = 3) = 0.$$

For example, from Bayes' rule,

$$Pr(\omega = 1) = \frac{1 * .5}{1 * .5 + .5(1 * (1-\epsilon) + 0 * \epsilon)} = \frac{1}{2-\epsilon}.$$

Expected payoffs are thus

$$u(\text{Attack}|\text{"No attack"}) = \frac{1}{2-\epsilon}0 + \frac{1-\epsilon}{2-\epsilon}\alpha, \text{ and}$$
$$u(\text{Don't attack}|\text{"No attack"}) = \frac{1}{2-\epsilon}(1-\epsilon-\epsilon\beta) + \frac{1-\epsilon}{2-\epsilon}.$$

Solving gives the expression in the Proposition, that the cooperative equilibrium is feasible when $\epsilon < \bar{\epsilon} = (2-\alpha)/(2-\alpha+\beta)$.

Proof of Proposition 3. Take "the strategy $\hat{z}$" to mean "attack if and only $z > \hat{z}$" (ignoring the knife-edge case). From definitions we have

$$G(z, +\infty) = S(z, +\infty) = \frac{2(1-\epsilon)}{2-\epsilon+\epsilon L(z)} \text{ and}$$
$$G(z, -\infty) = S(z, -\infty) = 0.$$

The strategy $\hat{z} = \infty$, or "do not attack for any signal," is an equilibrium only if for all signals

$z \in \mathbb{R}$,

$$\begin{aligned}(1+\beta)S(z,\infty) - \beta &\geq \alpha G(z,\infty) \\ (1+\beta)\frac{2(1-\epsilon)}{2-\epsilon+\epsilon L(z)} - \beta &\geq \alpha \frac{2(1-\epsilon)}{2-\epsilon+\epsilon L(z)}, \text{ or} \\ L(z) &\leq \frac{2(1-\epsilon)(1-\alpha)}{\beta\epsilon} - 1,\end{aligned}$$

which is the condition in the Proposition.

The strategy $\hat{z} = -\infty$, both attack for all signals, is an equilibrium if for all z

$$\begin{aligned}(1+\beta)S(z,-\infty) - \beta &\geq \alpha G(z,-\infty) \\ -\beta &\geq 0,\end{aligned}$$

which is true for all z so this in an equilibrium.

Last, $\hat{z} \in \mathbb{R}$ is an equilibrium if attacking is preferred to not attacking for any signal $z > \hat{z}$. So we have an equilibrium if for $z > \hat{z}$,

$$\begin{aligned}(1+\beta)S(z,\hat{z}) - \beta &< \alpha G(z,\hat{z}) \\ (1+\beta)(p_1(z)(1-\epsilon)F_0(\hat{z}) + p_2(z)) - \beta &< \alpha(p_1(z)(1-\epsilon)F_1(\hat{z}) + p_2(z)) \\ p_1(z)(1-\epsilon)[(1+\beta)F_0(\hat{z}) - \alpha F_1(\hat{z})] + p_2(z)[1+\beta-\alpha] &< \beta \\ 2\frac{1-\epsilon}{D}F_0(\hat{z})(1+\beta) - \alpha\left[\frac{(1-\epsilon)F_0(\hat{z})}{D} + \frac{(1-\epsilon)F_1(\hat{z})}{D}\right] &< \beta \\ 2(1-\epsilon)F_0(\hat{z})(1+\beta) - \alpha(1-\epsilon)(F_0(\hat{z}) + F_1(\hat{z})) &< \beta D.\end{aligned}$$

At $z = \hat{z}$, the two sides are equal (by definition of $\hat{z}$). βD is increasing in z because D increases with z, so the inequality holds for all $z > \hat{z}$. □

Proof of Proposition 4. Equation (3) results from algebra applied to find the z that solves

$$(1+\beta)S(z,\hat{z}) - \beta = \alpha G(z,\hat{z})$$

in terms of the likelihood ratio $L(z)$. Equation (4) takes further steps using the Normal distribution.

$z'_{br}(\hat{z}) < 1$ implies that the best reply to $\hat{z} + \delta$, where δ is a small number, differs from z^* by less in absolute value than $\hat{z} + \delta$. So a sequence will lead back to any z^* such that $z^* = z_{br}(z^*)$.

The highest z^* must have $z'_{br}(z^*) < 1$ because $R(\hat{z})$ approaches a finite limit, and thus eventually passes permanently below $L(z)$.

Further claims: For Normal F_a, $R(z)$ is unimodal and "upside down U" shaped. From taking the first derivative of R we have that it is increasing when (omitting the z arguments)

$$[(2-\alpha)f_0 - \alpha f_1]\left[\frac{1}{1-\epsilon} - F_0\right] + f_0[(2-\alpha)F_0 - \alpha F_1] > 0$$
$$[2-\alpha-\alpha L][\frac{1}{1-\epsilon} - F_0] + (2-\alpha)F_0 - \alpha F_1 > 0.$$

For Normal F_a this expression is positive for all z sufficiently small ($z \to -\infty$), so $R(z)$ must be increasing up to some point as we increase z from $-\infty$. For sufficiently large z, the expression is certainly negative since $L(z)$ gets very large; so for large enough z $R(z)$ is decreasing. If we can show that the expression has a negative first derivative for all z, then there can be only one z where it crosses zero, which would imply that $R(z)$ is unimodal. The derivative of the expression is

$$-\alpha L'(\frac{1}{1-\epsilon} - F_0) - f_0(2-\alpha-\alpha L) + (2-\alpha)f_0 - \alpha f_1$$

which reduces to $-\alpha L'\left(\frac{1}{1-\epsilon} - F_0\right)$ which is negative for all z. □

The characteristics of the Normal distribution used here are that its $L(z)$ is monotone increasing and goes from zero to positive infinity over the range of z. Any F_a that has the features, or in fact any L that gets "sufficiently large," will imply the same shape of $R(z)$.

□

Proof of Proposition 5. For α, β, and ϵ, this is immediate from observing that increasing each of these lowers $R(\hat{z})$ for given $\hat{z}$, and the fact that $R(z)$ cuts $L(z)$ from above at a stable equilibrium point.

[for σ, still working on this.]

...

Asymmetric case

z_1^{br} solves

$$L(z_1^{br}) = R(\hat{z}_2) \equiv \frac{(2-\alpha_1)F_0(\hat{z}_2) - \alpha_1 F_1(\hat{z}_2)}{\beta_1\left(\frac{1}{1-\epsilon_2} - F_0(\hat{z}_2)\right)} - 1.$$

z_2^{br} solves

$$L(z_2^{br}) = R(\hat{z}_1) \equiv \frac{(2-\alpha_2)F_0(\hat{z}_1) - \alpha_2 F_1(\hat{z}_1)}{\beta_2\left(\frac{1}{1-\epsilon_1} - F_0(\hat{z}_1)\right)} - 1.$$

Asymmetric case expected payoffs in equilibrium:

Let $F_{ab} = F_a(\hat{z}_b)$ for $a \in \{0,1\}$, $b \in \{1,2\}$ for states 1 and 2.

$$u_1^* = .5\left[(1-F_{01})[(1-\epsilon_2)F_{02} - \beta(1-(1-\epsilon_2)F_{02}] + \alpha(1-\epsilon_2)F_{01}F_{02}\right]$$
$$+ .5\left[(1-\epsilon_2)F_{02}[F_{01} + \alpha(1-F_{01})] - \beta(1-(1-\epsilon_2)F_{02})F_{11}\right].$$

$$u_2^* = .5\left[(1-F_{02})[(1-\epsilon_1)F_{01} - \beta(1-(1-\epsilon_1)F_{01}] + \alpha(1-\epsilon_1)F_{02}F_{01}\right]$$
$$+ .5\left[(1-\epsilon_1)F_{01}[F_{02} + \alpha(1-F_{02})] - \beta(1-(1-\epsilon_1)F_{01})F_{12}\right].$$

Accidents

A state's expected payoff for Attack is the same as in the "crazy type" interpretation, because a deliberate attack "overrides" any accidental launch. Writing p_1 for $p_1(z)$ and F_1 for $F_1(\hat{z})$),

$$u(\text{Attack}) = p_1\left[(1-\epsilon)(F_1\alpha + (1-F_1)0) + \epsilon 0\right] + p_2\alpha + p_3 0$$
$$= p_1(1-\epsilon)F_1\alpha + p_2\alpha = \alpha G(z,\hat{z}).$$

The expression for a state's expected payoff for Not Attack differs, however, because even if

one chooses to not attack, an accident could still occur. We have[26]

$$
\begin{aligned}
u(\text{Not Attack}) &= p_1\left[\epsilon\left[(1-\epsilon)(F_1\alpha+(1-F_1)0)\right]+\epsilon 0\right]+(1-\epsilon)\left[(1-\epsilon)\left[F_0+(1-F_0)(-\beta)\right]+\epsilon(-\beta)\right] \\
&\quad + p_2\left[(1-\epsilon)1+\epsilon\alpha\right]+p_3\left[(1-\epsilon)(-\beta)+\epsilon 0\right]. \\
&= \left[p_1(1-\epsilon)^2F_0+p_2(1-\epsilon)\right]*1-\beta\left[p_1\epsilon(1-\epsilon)+p_1(1-\epsilon)^2(1-F_0)+p_3(1-\epsilon)\right] \\
&\quad + \alpha\left[p_1\epsilon(1-\epsilon)F_1+p_2\epsilon\right] \\
&= (1-\epsilon)\left[p_1(1-\epsilon)F_0+p_2-\beta\left[p_1(\epsilon+(1-\epsilon)(1-F_0))+p_3\right]\right]+\epsilon\alpha\left[p_1(1-\epsilon)F_1+p_2\right] \\
&= (1-\epsilon)\left[p_1(1-\epsilon)F_0+p_2-\beta\left[1-(p_1(1-\epsilon)F_0-p_2)\right]\right]+\epsilon\alpha\left[p_1(1-\epsilon)F_1+p_2\right] \\
&= (1-\epsilon)\left[S(z,\hat{z})-\beta(1-S(z,\hat{z}))\right]+\epsilon\alpha G(z,\hat{z}).
\end{aligned}
$$

[26]Using $p_3 = 1 - p_1 - p_2$ in the third step.

References

Acemoglu, Daron and Alexander Wolitzky. 2014. "Cycles of conflict: An economic model." *American Economic Review* 104(4):1350–67.

Acemoglu, Daron and Matthew O Jackson. 2014. "History, expectations, and leadership in the evolution of social norms." *The Review of Economic Studies* 82(2):423–456.

Angeletos, George-Marios, Christian Hellwig and Alessandro Pavan. 2007. "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks." *Econometrica* 75(3):711–756.

Aveni, Thomas J. 2008. A Critical Analysis of Police Shootings under Ambiguous Circumstances. Technical report The Police Policy Studies Council.

Baliga, Sandeep, David O Lucca and Tomas Sjöström. 2011. "Domestic political survival and international conflict: is democracy good for peace?" *The Review of Economic Studies* 78(2):458–486.

Baliga, Sandeep and Tomas Sjöström. 2004. "Arms races and negotiations." *The Review of Economic Studies* 71(2):351–369.

Baliga, Sandeep and Tomas Sjöström. 2012. "The strategy of manipulating conflict." *American Economic Review* 102(6):2897–2922.

Boix, Carles and Milan W Svolik. 2013. "The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships." *The Journal of Politics* 75(2):300–316.

Bracken, Jerome and Martin Shubik. 1993. "Crisis stability games." *Naval Research Logistics (NRL)* 40(3):289–303.

Carlsson, Hans and Eric Van Damme. 1993. "Global games and equilibrium selection." *Econometrica: Journal of the Econometric Society* pp. 989–1018.

Chassang, Sylvain. 2010. "Fear of miscoordination and the robustness of cooperation in dynamic global games with exit." *Econometrica* 78(3):973–1006.

Dewan, Torun and David P Myatt. 2008. "The qualities of leadership: Direction, communication, and obfuscation." *American Political science review* 102(3):351–368.

Edmond, Chris. 2013. "Information manipulation, coordination, and regime change." *Review of Economic Studies* 80(4):1422–1458.

Ellison, Glenn. 1993. "Learning, local interaction, and coordination." *Econometrica: Journal of the Econometric Society* pp. 1047–1071.

Jackson, Van. 2019. *On the brink: Trump, Kim, and the threat of nuclear war*. Cambridge University Press.

Kandori, Michihiro, George J Mailath and Rafael Rob. 1993. "Learning, mutation, and long run equilibria in games." *Econometrica: Journal of the Econometric Society* pp. 29–56.

Kean, Thomas. 2011. *The 9/11 commission report: Final report of the national commission on terrorist attacks upon the United States*. Government Printing Office.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European Revolution of 1989." *World Politics* 44:7–48.

Little, Andrew T. 2016. "Communication technology and protest." *The Journal of Politics* 78(1):152–166.

Little, Andrew T. 2017. "Coordination, learning, and coups." *Journal of Conflict Resolution* 61(1):204–234.

Little, Andrew T et al. 2012. "Elections, fraud, and election monitoring in the shadow of revolution." *Quarterly Journal of Political Science* 7(3):249–283.

Morris, Stephen and Hyun-Song Shin. 2003. Global Games: Theory and Applications. In *Advances in Economics and Econometrics*, ed. Mathias Dewatripont, Lars Hansen and Stephen Turnovsky. Cambridge: Cambridge University Press.

Powell, Robert. 1989. "Crisis stability in the nuclear age." *American Political Science Review* 83(1):61–76.

Roessler, Philip. 2011. "The Enemy Within: Personal Rule, Coups, and Civil War in Africa." *World Politics* 63(2):300–346.

Roessler, Philip. 2017. *Ethnic Politics and State Power in Africa: The Logic of the Coup-Civil War Trap*. New York: Cambridge University Press.

Sagan, Scott Douglas. 1993. *The limits of safety: Organizations, accidents, and nuclear weapons*. Princeton University Press.

Savranskaya, Svetlana V. 2005. "New sources on the role of Soviet submarines in the Cuban Missile Crisis." *Journal of Strategic Studies* 28(2):233–259.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. New Haven: Yale University Press.

Schelling, Thomas C. 2006. *Micromotives and macrobehavior*. WW Norton & Company.

Wohlstetter, Roberta. 1962. *Pearl Harbor: warning and decision*. Stanford University Press.