The “Commitment Trap” Revisited: Experimental Evidence on Ambiguous Nuclear Threats

Michal Smetana*, Marek Vranka and Ondrej Rosendorf

Charles University, Faculty of Social Sciences, Smetanovo nabrezi 6, 110 00, Prague, Czech Republic
*Corresponding author. Email: smetana@fsv.cuni.cz

Abstract
In this paper, we provide an empirical test for the theoretical claim that ambiguous nuclear threats create a “commitment trap” for American leaders: when deterrence fails, supposedly they are more likely to order the use of nuclear weapons to avoid domestic audience costs for backing down. We designed an original survey experiment and fielded it to a sample of 1,000 U.S. citizens. We found no evidence of a commitment trap when ambiguous nuclear threats are made. Unlike explicit threats, ambiguous ones did not generate domestic disapproval when the leader backed down; the decision to employ nuclear weapons led to more public backlash for the leader than being caught bluffing; and the threats did not influence public preference for nuclear use across our scenarios. Our findings contribute to the scholarly literature on nuclear crisis bargaining and policy debates over the future of US declaratory policy.

Keywords: nuclear weapons; deterrence; calculated ambiguity; audience costs; weapons of mass destruction; survey experiment

Introduction
In his seminal International Security article, Sagan (2000) made a case against the use of nuclear threats to deter chemical and biological weapon attacks. To this end, the US declaratory policy has long been based on the principle of “calculated ambiguity” regarding the nature of its response. The policy allows American leaders to engage in deliberately vague messaging that implicitly entails the possibility of nuclear retaliation to unconventional attacks against the USA, its troops, or its allies (Lanoszka and Scherer 2017). Perhaps the most prominent real-world example of such ambiguous nuclear threats was the US attempt to deter Iraq from using...
chemical weapons during the First Gulf War: in 1991, President Bush sent Saddam Hussein a message that “[t]he United States will not tolerate the use of chemical or biological weapons [...] You and your country will pay a terrible price if you order unconscionable acts of this sort,” while Secretary of State Baker added that “American people will demand vengeance. And we have the means to exact it” (Buch and Sagan 2013).

According to Sagan, such ambiguous nuclear threats do not merely reflect a commitment to use nuclear weapons; they also create a commitment to do so. As a result, if deterrence fails, American leaders might end up being caught in a “commitment trap,” forcing them to employ nuclear weapons to avoid the reputational costs for backing down. In Sagan’s (2000, 87) words, “if [chemical weapons] or [biological weapons] are used despite such threats, the U.S. president would feel compelled to retaliate with nuclear weapons to maintain his or her international and domestic reputation for honoring commitments.” If this argument holds, calculated ambiguity increases the risk of nuclear use in crises. However, we lack empirical evidence that the relevant audiences indeed perceive ambiguous threats according to Sagan’s theoretical assumptions.

In this paper, we examine whether the commitment trap argument holds vis-à-vis the American public. We fielded an original survey experiment in the USA to test several hypotheses theoretically grounded in “audience costs” literature (Fearon 1994; Tomz 2007; Levendusky and Horowitz 2012; Kertzer and Brutger 2016). Using vignettes and fictional social media posts to describe the development of a crisis between the USA and North Korea, we experimentally manipulated the formulation of the US President’s deterrence threat and the subsequent US response when the threat failed to deter North Korea’s chemical attack.

Our results show little evidence for a commitment trap with respect to ambiguous nuclear threats. Unlike explicit nuclear threats, ambiguous ones did not generate domestic disapproval in the case of backing down from nuclear use. Moreover, the decision to order nuclear strikes led to more public backlash than being caught bluffing. Finally, neither ambiguous nor explicit nuclear threats influenced public preference for nuclear use.

In the following sections, we (1) present our theoretical framework, (2) formulate our hypotheses, (3) introduce our experimental design, (4) present the results, and (5) discuss the implications of our findings.

Theoretical framework

Our approach to the study of the “commitment trap” is theoretically grounded in the “audience costs” literature. Fearon (1994) originally coined the concept of audience costs to explain why democracies are able to signal military threats in crisis bargaining more credibly than authoritarian states. He proposed that leaders in democracies that escalate and then back down suffer a loss of popularity at home. The micro-mechanism behind this effect is the public preference for consistency driven by reputational or normative concerns (Tomz 2007, 833–36). Publicly issued threats supposedly “tie hands” of leaders, which makes signaling inherently more
credible given the *ex-post* audience costs that would be generated if the leaders do not follow through with their commitment (Fearon 1997).¹

Let us follow Kertzer and Brutger (2016, p. 237) by visualizing this dynamic through two types of graphical representations. In Figure 1, we show a nuclear crisis bargaining game tree, with nodes indicating the player and the decision-making point in time.² In node 1, Player 1 is deciding whether to issue a nuclear threat ($\alpha_2$) or not ($\alpha_1$). Player 2 decides whether to attack or not in nodes 2a and 2b, respectively. If she does, Player 1 faces an option of whether to order the use of nuclear weapons ($\beta_1$) or not ($\beta_2$) in corresponding nodes 1a and 1b. The letter $\mu$ then shows the pay-off for Player 1, which corresponds to the public approval as the game resolves. Table 1 shows this decision-making structure as a $2 \times 2$ experimental notation.

If ambiguous nuclear threats indeed create a commitment trap for the US President, not following through with the threat should generate domestic audience costs. As other scholars have found in survey experiments on foreign interventions (Kertzer and Brutger 2016; Levendusky and Horowitz 2012; Tomz 2007), the level of public approval should be lower when the leader backs down from the threat ($\mu_{22}$) than in a corresponding scenario where the leader did not issue the threat in the first place ($\mu_{12}$). When all other aspects of the scenario are equal, the difference in

---

¹Kertzer and Brutger (2016) later demonstrated that audience costs can be further disaggregated into “inconsistency costs” and the “belligerence costs.” Due to space constraints, we only report the “composite” audience costs as Tomz (2007), Levendusky and Horowitz (2012), Lin-Greenberg (2019), and others.

²The logic is applicable to both ambiguous and explicit nuclear threats, and we investigate them individually in the subsequent analyses.
approval between the two outcomes ($\mu_{12} - \mu_{22}$) is the absolute domestic audience cost paid by the leader.

It is worth noting that our approach partially deviates from the traditional audience costs experiments by operationalizing the “back down” option $\beta_2$ as a non-nuclear yet still military response rather than “doing nothing.” The main reason was to capture the actual policy dilemma discussed by both Sagan (2000, pp. 112–5) and his critics (Martin and Sagan 2001, p. 193), which is about choosing between a nuclear or conventional military response rather than a nuclear or nonmilitary response. A scenario where the US president does nothing in response to a chemical strike against US troops and allies is arguably unrealistic; in fact, there is a recent precedent of US military strikes in response to Syria’s chemical use even though neither American troops nor allies came to harm, with US President scoring approval points across the political spectrum for ordering these strikes (Doucet 2018).

Our approach may, therefore, resemble the work of Lin-Greenberg (2019), who found that leaders could reduce audience costs for not following through with their threat by “backing up” to a less hawkish policy (e.g., air strikes or economic sanctions rather than a full invasion in the traditional audience costs setup). However, we do not know whether “backing up” is acceptable for the US public even when the chemical “taboo” is violated (cf. Bentley 2014) and American lives are lost, which is a considerably more extreme case than a foreign intervention described in the traditional audience costs scenarios. Moreover, Lin-Greenberg only investigated policy substitution strategies for explicit threats, and it remains to be seen whether “backing up” after ambiguous threats also merely reduces audience costs or eliminates them altogether.

Hypotheses
To test the audience costs logic for ambiguous nuclear threats, we formulate the following hypothesis:

**H1:** Leaders suffer domestic disapproval when they issue ambiguous nuclear threats and then back down from using nuclear weapons when deterrence fails.

If ambiguous nuclear threats do not generate audience costs, there are two explanations for why that might be the case. It is conceivable that the vague formulation of the threat does not create the commitment in the same way an explicit threat would. However, it is also possible that the public simply does not punish the leader for not following through with nuclear use, its specific formulation notwithstanding. If the former applies, we expect that the audience costs for the explicit threat will be higher than for the ambiguous threat ($\mu_{12} - \mu_{22}^{\text{EXP}} > \mu_{12} - \mu_{22}^{\text{AMB}}$). If
the latter applies, we expect that both threats generate comparable audience costs \((\mu_{12} - \mu_{22}^{\text{EXP}} = \mu_{12} - \mu_{22}^{\text{AMB}})\). This leads us to two competing hypotheses:

**H2a:** Explicit threats generate larger audience costs than ambiguous threats.

**H2b:** There is no statistically significant difference between audience costs generated by explicit and ambiguous threats.

Under the commitment trap logic, the leaders perceive that they must use nuclear weapons to avoid public resentment for backing out when deterrence fails. The public should, therefore, disapprove more of the leaders’ empty threats than of the actual nuclear use \((\mu_{21} > \mu_{22})\). Otherwise, it would be more beneficial for the leaders to renge on their commitment than undertake an action that generates more public backlash than being caught bluffing.

**H3a:** When deterrence fails following the leaders’ ambiguous nuclear threat, the public is less likely to approve of their handling of the crisis if they do not follow through with nuclear use.

**H3b:** When deterrence fails following the leaders’ ambiguous nuclear threat, the public is less likely to approve of their handling of the crisis if they follow through with nuclear use.

Another piece of evidence for the commitment trap would be a higher public preference for nuclear use following the leader’s ambiguous threat. Earlier studies have found that the public preference for nuclear use is subject to consequentialist reasoning (Dill, Sagan, and Valentino 2022; Press, Sagan, and Valentino 2013). In our study, the consequentialist reasoning could be related to the concern that not following through with nuclear use after an ambiguous nuclear threat could negatively impact the country’s reputation and the credibility of US coercive signaling in future crises.

**H4:** When deterrence fails, the public is more likely to prefer the use of nuclear weapons if their leader issued an ambiguous nuclear threat.

**Experimental design**

We designed an original survey experiment with \(3 \times 2\) conditions and fielded it to a sample of 1,001 adult Americans through the Prolific online platform. We used quotas on gender and political identification to get a more balanced and representative sample. Additionally, we obtained individual-level data on age, income, and education to be used as control variables in our analyses.\(^3\)

\(^3\)See Appendix 1 for the sociodemographic composition, Appendix 2 for survey items, and Appendix 6 for ethical considerations. In Appendices 11–15, we report the results of our follow-up experiments conducted during the peer-review process.
In the survey vignette, we described a development of a crisis involving North Korea, the USA, and US allies in the region. Our participants were randomly assigned to three treatment groups, where each read one version of a tweet posted by a new US President Smith. In the control group ($\alpha_1$), the President announced that the government received intelligence on an impending North Korean chemical attack (see Figure 2). In the “ambiguous threat” treatment ($\alpha_2^{\text{AMB}}$), the President added that “if North Korea was foolish enough to use chemical weapons, our response will be absolutely overwhelming and devastating, and all military options will be on the table!” In the “explicit threat” treatment ($\alpha_2^{\text{EXP}}$), the President stated that the United States would strike back with “our powerful nuclear arsenal” in the event of a chemical attack.

Next, we displayed information that 2 days after this announcement, North Korea used chemical weapons against the Japanese island of Okinawa, killing 1,400 Japanese civilians and 650 US troops stationed at the local US base. President Smith ordered the US armed forces to prepare a military response. We then randomly displayed one version of a tweet that described the nature (nuclear $\beta_1$ or non-nuclear $\beta_2$) of this military response (see Figure 3). Next, we asked about the approval or disapproval of the President’s handling of the crisis on a seven-point scale ($H_1$–$H_3$) and about the preference for using nuclear weapons in the scenario on a four-point scale ($H_4$). We concluded with a debrief (see Carpenter, Montgomery, and Nylen 2021).

---

4See Appendix 9 for a discussion of the advantages and disadvantages of using North Korea in our scenario.

5We fielded an additional experiment using an alternative wording of $\alpha_1$. The results (see Appendix 11) show that our original approach was more conservative and indeed a tougher test for an argument advanced in this paper.

6We used the words of the former US Secretary of Defense William Perry, which Sagan (2000, p. 85) highlights as an example of an ambiguous nuclear threat.
Results

First, we conducted an ordinal logistic regression with the approval of the President’s handling of the crisis as an outcome variable, treatment $\alpha$ as a predictor, and gender, age, income, education, and partisanship as control variables. We kept the nature of the military response constant, i.e., non-nuclear ($\beta_1$). As we show in Figure 4, there was no statistically significant difference ($p = 0.627$) between the

Figure 3.
Treatments $\beta_1$ and $\beta_2$.

Figure 4.
Ordinal logistic regression estimates. $N = 476$. 95% CI. Variables whose intervals overlap with the vertical line are statistically indistinguishable from 0. Positive coefficients correspond to a higher level of approval. Model 1 shows the effects without the inclusion of control variables and Model 2 with control variables included.
approval in the control group ($\alpha_1$) and ambiguous threat group ($\alpha_2^{AMB}$). We, therefore, reject hypothesis H1 that leaders suffer domestic disapproval when they make ambiguous nuclear threats and then back down from using nuclear weapons when deterrence fails.

However, Figure 4 also shows that approval was significantly lower ($p < 0.001$) in the explicit nuclear threat group ($\alpha_2^{EXP}$) than in the ambiguous threat group ($\alpha_1^{AMB}$). The absolute audience costs for the explicit threat are higher than the audience costs for the ambiguous threat ($\mu_{12} - \mu_{22}^{EXP} > \mu_{12} - \mu_{22}^{AMB}$). As such, we gained empirical support for hypothesis $H2a$ that explicit nuclear threats generate larger audience costs than ambiguous ones, and we reject the competing $H2b$ that these costs are statistically indistinguishable.

Next, we examined whether the public disapproves more of the leader’s empty threats than of the actual use of nuclear weapons (i.e., if $\mu_{21} > \mu_{22}$). Figure 5 shows the results for participants in the ambiguous treatment ($\alpha_2^{AMB}$) with approval as an outcome variable, response $\beta$ as a predictor, and sociodemographic characteristics as control variables. The nuclear response was negatively associated with approval ($p < 0.001$). As such, we reject $H3a$ and gain support for $H3b$ that the public

7See Appendix 3 for detailed analysis. For all hypotheses, we also report observed effects in percentages of selected responses (Appendix 8) and additional analyses after excluding participants who failed a manipulation check (Appendix 10).
disapproval is higher if the leaders employ nuclear weapons after an ambiguous nuclear threat than if they do not.\footnote{See Appendix 4 for detailed analysis. For a corresponding analysis in the $\alpha_{\text{EXP}}$ (“explicit threat”) group, see Appendix 7.}

Next, we investigated whether nuclear threats make our participants more likely to prefer the use of nuclear weapons. We conducted a logistic regression with preference as an outcome variable, treatment $\alpha$ as a predictor, and the response $\beta$, $\alpha \ast \beta$ interaction, and sociodemographic characteristics as control variables. As we show in Figure 6, there was no statistically significant association between preference for nuclear use and the ambiguous (or explicit) nuclear threat.\footnote{See Appendix 5 for detailed analysis.} Thus, we reject $H_4$ that the public is more likely to prefer the use of nuclear weapons if their leader issued an ambiguous nuclear threat beforehand.

Since we found that, unlike explicit threats, ambiguous ones did not generate any audience costs, we fielded an additional experiment to see why that might be the case. One possible explanation is that the public is not attentive to the implicit hint at nuclear use in ambiguous messaging (so it does not find the non-nuclear response inconsistent). We, therefore, examined the baseline expectations of the likelihood of different policy responses following $\alpha_{\text{AMB}}$ and $\alpha_{\text{EXP}}$ (due to space constraints, we discuss the design and full results in Appendix 13). In Figure 7, we show that
although our respondents saw nuclear use as slightly more likely in $\alpha^\text{EXP}$, the difference was not statistically significant ($p = 0.497$). This suggests that the (non-)effects found in our original experiment were not caused by the public ignorance of implicit reference to nuclear weapons in leaders’ ambiguous threats.

Discussion and conclusions

Our study found no evidence for the “commitment trap” in the context of ambiguous nuclear threats, questioning one of the key assumptions of Sagan’s (2000) seminal work on US declaratory policy. The fictional leader in our study did not pay domestic audience costs when he backed down from using nuclear weapons. However, when we exposed the participants to the explicit nuclear threat treatment, the approval decreased significantly once the leader reneged on his commitment. We also demonstrated that the US public is attentive to the implicit hints at nuclear use in ambiguous threats. As such, it probably does not see them as binding as explicit ones$^{10}$ and, therefore, is less likely to punish the President for not following through with nuclear use when deterrence fails. These findings give credence to

$^{10}$Our follow-up experiment on the perceived impact on the US credibility further supports this claim (see Appendix 15).
Susan Martin’s response to Sagan that a conventional response following an ambiguous nuclear threat would not impact the US reputation, as “negative reputational effects follow from the failure to carry out the threatened punishment, not from the failure to carry out the threatened punishment by a particular means” (Martin and Sagan 2001, p. 193).

Beyond addressing the original “commitment trap” argument, our work contributes to the burgeoning audience costs scholarship (Kertzer and Brutger 2016; Levendusky and Horowitz 2012; Lin-Greenberg 2019; Tomz 2007) by investigating pertinent scenarios involving weapons of mass destruction and exploring differences in public perceptions of ambiguous and explicit threats. As argued by Snyder and Borghard (2011), the latter is an important distinction for the audience costs theory, yet has rarely been tested in audience costs experiments. Our finding that the US public preference regarding nuclear use remains constant irrespective of leaders’ threats also adds to the recent wave of “nuclear taboo” experiments (Allison, Herzog, and Ko 2022; Bowen, Goldfien, and Graham 2023; Dill, Sagan, and Valentino 2022; Horschig 2022; Koch and Wells 2021; Press, Sagan, and Valentino 2013; Rathbun and Stein 2020; Sagan and Valentino 2017; Smetana and Vranka 2021; Smetana and Onderco 2023; Smetana, Vranka, and Rosendorf 2023; Sukin 2020).

Our experiment provides pertinent insights for policy debates over the US declaratory policy (Fetter and Wolfsthal 2018; Gerson 2010; Holdren 2020; Panda and Narang 2021; Roberts 2019; Sagan 2009). It appears that making ambiguous nuclear threats is a relatively cost-free strategy for US leaders: when deterrence fails, they can resort to a non-nuclear response without being punished for inconsistency by the domestic constituency. On the other hand, this is precisely what makes ambiguous threats less effective than explicit ones; as Fearon (1994) demonstrated, having the possibility of reneging on one’s commitment without suffering the loss of political points makes coercive signaling inherently less credible.

Yet, we must stress that while public attitudes represent an important piece of the “commitment trap” puzzle, it is not the whole story. There are other relevant audiences that American leaders consider. Notably, the formulation of US declaratory policy and the corresponding strategic messaging have always been heavily informed by the views of US allies in Europe and Asia (Horovitz et al. 2021; Perkovich and Vaddi 2021; Smetana 2018). Moreover, it is reasonable to assume that US enemies pay close attention to the credibility of US nuclear threats in their strategic calculations. Overall, we need new experimental evidence on how the “commitment trap” logic plays out vis-à-vis different kinds of international audiences.

Finally, we must stay open to the theoretical possibility that even if ambiguous threats do not generate domestic audience costs, US leaders believe that they do and make strategic decisions in accordance with these beliefs. If this argument holds, ambiguous threats could entrap US leaders exactly as Sagan originally proposed, actual public attitudes notwithstanding. Future studies should, therefore, make use of corresponding elite surveys and survey experiments (see, e.g., Dietrich, Hardt, and Swedlund 2021; Kertzer and Renshon 2022; Saunders 2022; Smetana...
and Onderco (2022) to find out more about the leaders’ perceptions of the “commitment trap.”

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/XPS.2023.8

Data availability statement. Support for this research was provided by the Charles University’s PRIMUS program (Award PRIMUS/22/HUM/005). The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/Q91W2O.

Acknowledgments. We thank Scott Sagan, Lauren Sukin, JEPS editorial team, and two anonymous reviewers for their valuable feedback. We also gratefully acknowledge funding from the Charles University’s program PRIMUS/22/HUM/005 (Experimental Lab for International Security Studies – ELISS).

Conflicts of interest. The authors declare no conflicts of interest.

Ethics statement. The data collection for this research study adhered to all relevant aspects of APSA’s Principles and Guidance for Human Subjects Research and to the general data collection approach used in the Charles University’s research project “Experimental Lab for International Security Studies (ELISS)” that was approved without reservations by the Commission for Ethics in Research of the Faculty of Social Sciences, Charles University (submission #53).

References


https://doi.org/10.1017/XPS.2023.8 Published online by Cambridge University Press


Cite this article: Smetana M, Vranka M, and Rosendorf O. The “Commitment Trap” Revisited: Experimental Evidence on Ambiguous Nuclear Threats. Journal of Experimental Political Science. https://doi.org/10.1017/XPS.2023.8