

Let's Get Reel: The Limits of AI Video for Experimental International Relations

Michal Smetana, Marek Vranka, Lauren Sukin, and Ondrej Rosendorf

August 31, 2026

Note: This is a preprint of an academic paper that has not yet been peer-reviewed.
The paper will be updated continuously.

Authors:

(1) Michal Smetana is an Associate Professor of International Studies at the Faculty of Social Sciences, Charles University, Director of the Peace Research Center Prague, and the Principal Investigator of the European Research Council (ERC) project MICROCODE.

E-mail: smetana@fsv.cuni.cz

Phone number: +420 723065858

Address: Charles University, Faculty of Social Sciences, U Krize 8, 158 00, Prague, Czech Republic
(*corresponding author*)

(2) Marek Vranka is a researcher at Peace Research Center Prague, a lecturer at the Faculty of Social Sciences, Charles University, and leader of the Prague Experimental Laboratory for Social Sciences.

E-mail: marek.vranka@fsv.cuni.cz

(3) Lauren Sukin is the John G. Winant Associate Professor in U.S. Foreign Policy at Nuffield College & Department of Politics and International Relations, University of Oxford, and a researcher at the Peace Research Center Prague at the Faculty of Social Sciences, Charles University.

E-mail: Lauren.sukin@politics.ox.ac.uk

(4) Ondrej Rosendorf is a postdoctoral researcher at the Peace Research Center Prague, Faculty of Social Sciences, Charles University, and the Institute for Peace Research and Security Policy at the University of Hamburg.

E-mail: ondrej.rosendorf@fsv.cuni.cz

Abstract

Generative AI tools promise to make survey experiments in International Relations (IR) more immersive by delivering treatments through realistic, artificially generated videos. But would these novel designs offer real methodological gains over traditional text-based vignettes? To address this question, we conducted a preregistered experiment with a quota-representative sample of U.S. citizens ($N = 2,243$). Participants responded to a hypothetical scenario of a Russian attack on Latvia, a U.S. ally on NATO's eastern flank, indicating their support for different policy responses. The scenario was presented in one of four randomly assigned "communication modalities": an AI-generated video, long text with AI-generated images, plain long text, and a short-text vignette commonly used in IR survey research. Contrary to our expectations, the AI-generated video did not elicit the strongest affective responses or the largest shifts in policy preferences; its effects were statistically indistinguishable from plain long-form text and were even marginally weaker than the text-with-images treatment. The video's one advantage was participants' self-reported satisfaction with the survey, in which it outscored the plain-text conditions but not the text-with-images vignette. However, all three long-form modalities outperformed the short vignette across every outcome measure except attentiveness to the experimental manipulation. Overall, our study found no evidence that AI-generated videos provide IR experimentalists with substantial methodological benefits. Instead, immersive designs that combine engaging text-based descriptions with AI-generated images may offer a cost-effective way to enhance the mundane realism and ecological validity of treatments in IR survey experiments.

Keywords

survey experiment; experimental IR; research design; communication modality; generative AI; foreign policy attitudes; ecological validity

Introduction

For much of its history, the field of International Relations (IR) has mostly shied away from using experimental methods. To many, studying how states, leaders, and international organizations behave seemed irreconcilable with experimental designs commonly used in social psychology, sociology, and some subfields of political science and economics (McDermott 2002, 2011). However, the turn towards behavioral approaches (Hafner-Burton et al. 2017; Davis and McDermott 2021) and the greater affordability of online polling have led to a surge in IR survey experiments over the last few decades (Hyde 2015; Kertzer 2021). The particular focus—and, arguably, the largest contribution—of these studies has been to investigate the microfoundations of macro-level phenomena in world politics, testing many of the key assumptions in our field empirically through survey experiments with both publics and elite decision-makers (Kertzer 2017; Dietrich et al. 2021; Kertzer and Renshon 2022).

This is not to say that the employment of survey experiments in IR has been without critique. A relatively common one concerns the use of short text-based vignettes describing hypothetical situations that do not resemble how individuals consume information about events in international politics (Barabas and Jerit 2010; Egami and Hartman 2023). The potential lack of mundane realism and concern about ecological validity led some IR scholars to employ elaborate mock news stories (Press et al. 2013; Dill et al. 2022; Wayne 2023; Schwartz 2024; Logan 2025), fake social media posts (Harris and Lin-Greenberg 2024; Smetana et al. 2024), or cellphone emergency notifications (Allison et al. 2022) as experimental stimuli, with scholars still examining and debating the advantages and disadvantages of these approaches (Kreps and Roblin 2019; Brutger et al. 2022; Harris and Lin-Greenberg 2025). Today, the advent of new generative AI tools offers an intriguing promise: designing even more immersive survey experiments that administer treatments through realistic videos generated by artificial intelligence (Becker and Laycock 2023; Schwitter 2025). But would such an approach ultimately provide any substantive methodological benefits?

So far, no studies in our field have tested the relative advantages and disadvantages of AI-generated video designs compared to text-based or text-and-images vignettes commonly used by IR scholars. In this article, we aim to address this gap through an original survey experiment that tests the relative effectiveness of these different “communication modalities” within the broader study of public attitudes toward collective defense. In a preregistered online study of American citizens ($N = 2,243$), we presented participants with a scenario of Russian aggression against Latvia, a U.S. ally geographically positioned on NATO’s eastern flank. We randomly assigned participants to 2×4 treatment conditions based on the type of Russian aggression (a conventional military invasion or a covert “hybrid” operation) and the communication modality for the scenario displayed in the survey (AI-generated video, long text with AI-generated images, long text, and short text-based vignette). We then examined the relative performance of each communication modality in terms of participants’ affective response to the treatment, policy preference in the given scenario, attentiveness, and overall satisfaction with taking the survey.

We found that, contrary to our expectations, it was not the AI-generated video design that elicited the strongest affective response. Although this modality performed better than the traditional short-text vignette, its effect was statistically indistinguishable from a long-text description of the scenario and was even marginally weaker than that of the text-with-images condition. The level of public support for a military response to defend Latvia was comparable across all three long-version vignette treatments and significantly weaker in the short-version treatment. Respondent attentiveness, measured by successful completion of the manipulation check, was generally high, with the highest level in the short vignette version, suggesting that the brief text may have made the key information more salient for participants in our experiment. We found the strongest evidence in favor of using the AI-generated video for participants’ satisfaction with the survey, with the video seen as particularly engaging and, together with text-with-images, the most positively evaluated communication modality.

The results of our study suggest that we should not be overly optimistic about the AI-generated videos significantly improving the effectiveness of IR survey experiments. What appears clear from our findings is that more elaborate, longer, and more immersive vignettes that aim to engage participants and draw them more “into” the scenario can elicit significantly different emotional responses and, in turn, yield different results with respect to the outcome measures of interest. The clear “winner” among the communication modalities we tested in our experiment is exactly this type of long, engaging text accompanied by realistic AI-generated images. Given that this approach also requires significantly fewer resources than generating realistic videos for individual treatment arms, our tentative recommendation to our colleagues is that, in the absence of further evidence, the immersive text-with-images designs represent the optimal communication modality for IR survey experiments when realism and ecological validity are key design considerations.

We proceed as follows. First, we review the use of various communication modalities in IR survey experiments and the limited attempts in our field to test their relative effectiveness. Second, we discuss our theoretical expectations and hypotheses regarding the use of AI-generated visual survey designs. Third, we introduce the design of our experiment. Fourth, we summarize the results of our preregistered analyses. We conclude by discussing the implications of our findings for IR scholarship.

Communication Modalities in IR Survey Experiments

The typical way to conduct survey experiments in IR has always been to administer brief, text-based vignettes describing variations of a hypothetical scenario in world politics. For example, Tomz (2007), one of the most-cited survey experimental studies in IR, is based on a two-paragraph, fairly abstract scenario that describes how an unnamed state “sent its military to take over a neighboring country” and how the U.S. president responded. The participants in the study were then asked whether they approved or disapproved of the president’s handling of the situation.

To increase the mundane realism of their experiments, some IR scholars have opted for more elaborate communication modalities, such as “mock news” vignettes that resemble actual newspaper articles in both writing style and graphical representation. Perhaps the most typical example of this approach is the “Atomic Aversion” study by Press, Sagan, and Valentino (2013), who exposed their participants to a fictional newspaper story about military strikes against Al Qaeda’s facility (see Dill et al. 2022; Wayne 2023; Schwartz 2024; Logan 2025 for a similar approach). Other scholars simulated the public communication of elite decision-makers through social media mockups, such as short-form “tweets” from official government accounts (Harris and Lin-Greenberg 2024; Smetana et al. 2024). Allison et al. (2022) used a mock emergency alert that mimics the notifications displayed on mobile phones (see Figure 1 for these three types of graphical vignettes).

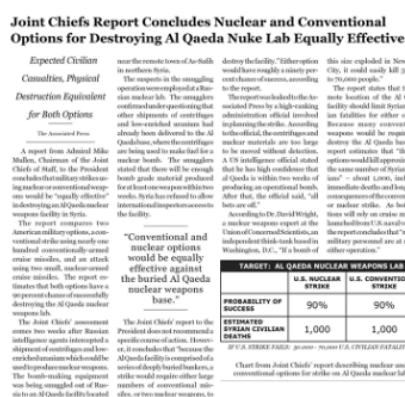


Figure 1. Examples of contextually realistic graphics in IR survey experiments. From left: news story (Press et al. 2013), social media post (Harris and Lin-Greenberg 2024), and emergency alert (Allison et al. 2022).

The jury is still out on the extent to which these approaches provide substantial methodological benefits over the more prevalent short-form text-based vignettes. A study by Kreps and Roblin (2019) found few differences in performance between plain-text and mock news, although the latter mitigated some confounding effects of short, abstract vignettes. Brutger et al. (2022) found no substantive changes in results from studies that used more abstract or more detailed experimental stimuli, but they did find that including more contextual information in the vignette ultimately dampened the effect of the experimental treatment. Harris and Lin-Greenberg (2025) tested contextually realistic graphics and found that their employment did not lead to substantially different outcomes compared with purely text-based vignettes.

Beyond the use of illustrative images and graphic elements, adjacent fields have also recently started employing “deepfakes” (Vaccari and Chadwick 2020; Dobber et al. 2021; Hameleers et al. 2022; Barari et al. 2025; Dan 2025; Weikmann et al. 2025; Hameleers et al. 2026) and AI-edited video footage (Wu et al. 2026) in their survey experiments. In IR, however, we have not yet observed scholars leveraging the AI revolution and employing AI-generated videos in experimental treatments. In fact, even the use of non-AI video formats has been extremely rare in IR so far, with Green-Riley et al. (2022)’s use of existing footage of the violent repression of public protests representing a rare example. Given that video (particularly distributed via social media platforms) has become a dominant form of news consumption (Newman et al. 2025), using AI-generated videos as a potentially engaging and realistic format for depicting international events in IR survey experiments appears to be a promising avenue. Yet, we still lack empirical evidence on whether video use would provide IR researchers with substantive benefits over existing approaches—an important gap our study aims to fill.

Theoretical Expectations and Hypotheses

Existing research on communication modalities—the form in which messages are communicated—already provides us with arguments for why and how AI-generated video could potentially perform differently in IR survey experiments than vignettes based on plain text or text accompanied by images. Early studies in this area found that some types of messages could be more persuasive in video form than in audio or plain text (Chaiken and Eagly 1976), although others showed that the type of medium can interact with the source of the message (Andreoli and Worchel 1978) and that the persuasive advantage can be relatively small in the context of political videos (Wittenberg et al. 2021). Across different situational contexts, studies have shown that video can elicit stronger emotional responses than text (Fishfader et al. 1996; Glasford 2013). Others have demonstrated the potential of the video modality for political fact-checking (Young et al. 2018) and for communicating scientific consensus (Goldberg et al. 2019). Through testing both traditional and virtual reality video vignettes against text, van Zelderen et al. (2025) found that participants generally preferred more immersive designs and that these designs can be useful for building more ecologically valid surveys (cf. Aguinis and Bradley 2014).

The revolution in generative AI tools has recently sparked debate over their potential to design realistic images and videos for survey experiments (Davidson 2024; Schwitter 2025). Eberl et al. (2022) demonstrated that, even without specialized knowledge, social scientists can generate convincing “deepfake” videos and use them effectively in experimental manipulations. However, while the deepfakes of political messaging have been employed quite extensively in survey designs over the past few years (Vaccari and Chadwick 2020; Dobber et al. 2021; Hameleers et al. 2022; Barari et al. 2025; Dan 2025; Weikmann et al. 2025; Hameleers et al. 2026), they primarily involve deepfake techniques such as face-swapping or voice-cloning rather than natively generated text-to-video or image-to-video. Overall, we have not yet observed attempts to use the new generative AI tools to design videos depicting complex situational developments relevant to IR scholars. And

that also means that, so far, we do not know whether these novel tools could provide tangible benefits in experimental studies of phenomena relevant to our field.

Our study aims to leverage these new text-to-video and image-to-video AI tools to design an original audiovisual vignette depicting a fictional international crisis that concerns an ally of our participants' country. As in other studies on the microfoundations of collective defense (Tomz and Weeks 2021; Tomz et al. 2023; Gomez et al. 2025; Smetana et al. 2026), participants face a dilemma about whether to support defending the victim of aggression under the given treatment. Although we also include some novel contextual manipulations in our experiment—such as varying between a conventional military invasion and a “hybrid” operation—the key focus of our study is on manipulating and testing the effectiveness of the communication modality: AI-generated video, long text accompanied by AI-generated images, plain long text, and a short survey vignette of the type most commonly used in IR survey research.

Building on the existing scholarly literature, we formulated *a priori* theoretical expectations about the benefits of an AI-generated video vignette relative to text-only and text-with-images modalities. Given the greater immersive potential of the video format, we expected that modality to generate a stronger affective response to the international crisis (cf. Glasford 2013), which would then lead to shifts in policy preferences (in our study, an increased support for defending the ally; see Wayne 2023 for a similar causal reasoning). We also expected that the participants would find this novel format significantly more engaging (cf. van Zelder et al. 2025), resulting in higher levels of attentiveness and satisfaction with completing the survey. We preregistered¹ the following formal hypotheses prior to data collection:

¹ We provide the preregistration protocol and the detailed discussion of the deviations from the preregistered sampling and analytical plans in Appendix 1. Note that although the hypotheses are formulated identically to those in the preregistration, we renumbered them so that they correspond to the survey flow discussed in the next section.

- ▽ **Affective response (H1):** *Participants who watch an AI-generated video vignette exhibit a stronger affective response to the treatment than participants who read a text-only vignette (H1a) or a vignette combining text and AI-generated images (H1b).*
- ▽ **Support for military engagement (H2):** *Participants who watch an AI-generated video vignette will be more supportive of direct military engagement than participants who read a text-only vignette (H2a) or a vignette combining text and AI-generated images (H2b).*
- ▽ **Attention (H3):** *Participants who watch an AI-generated video vignette are more attentive to the treatment than participants who read a text-only vignette (H3a) or a vignette combining text and AI-generated images (H3b).*
- ▽ **Satisfaction (H4):** *Participants who watch an AI-generated video vignette will evaluate the survey more positively than participants who read a text-only vignette (H4a) or a vignette combining text and AI-generated images (H4b).*

Experimental Design

To test our theoretical expectations on the potential benefits of using AI-generated videos in experimental IR, we designed and preregistered an original online survey experiment.² Figure 2 summarizes the logic of our experimental design and the survey flow, which we elaborate on in this section.

After consenting to the study and responding to a battery of sociodemographic questions, the participants were randomly assigned to four conditions based on four different communication modalities: an AI-generated video (treatment A), long text accompanied by AI-generated images (treatment B), plain long text (treatment C), and a short survey vignette (treatment D).

² See Appendix 2 for the full survey instrument.

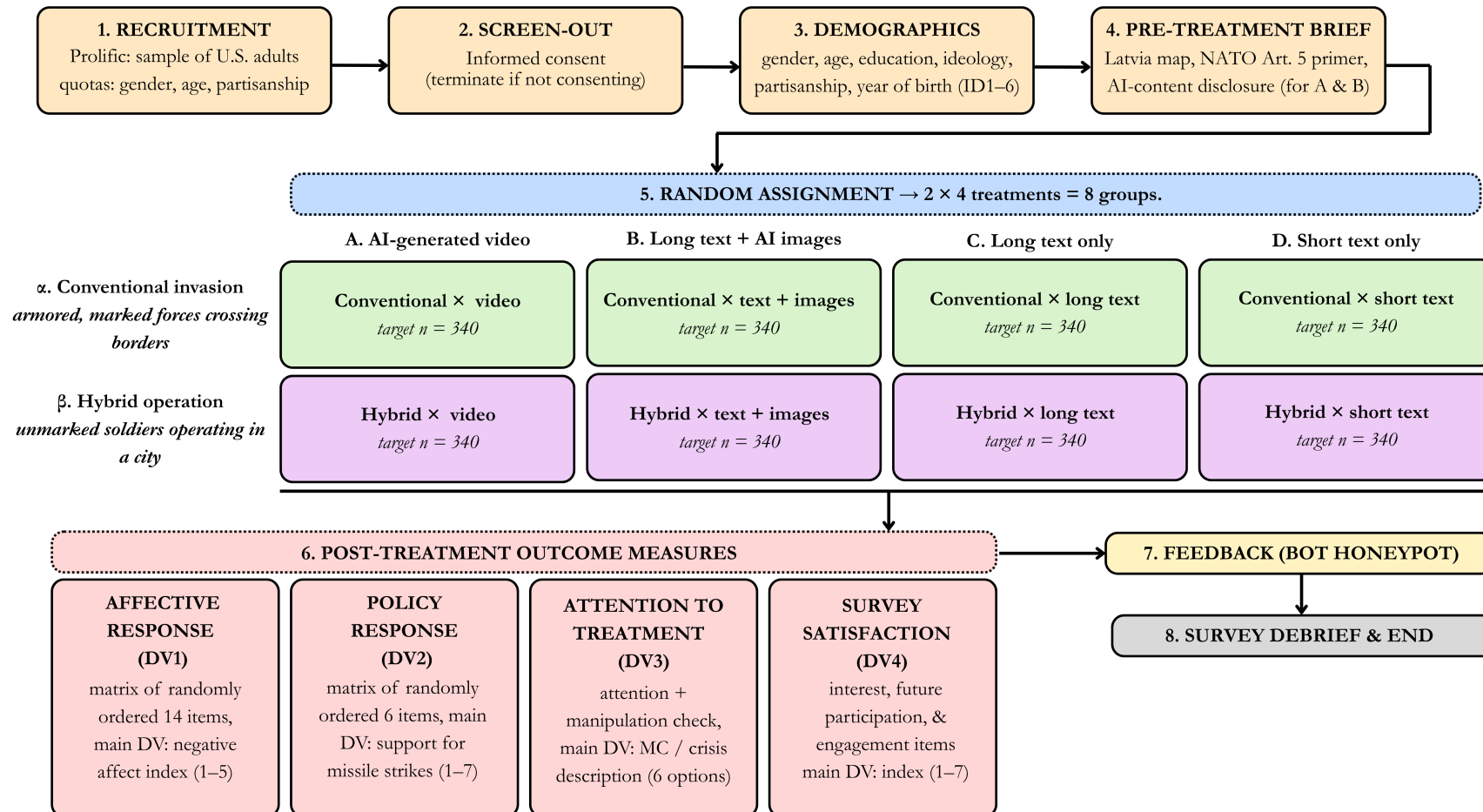


Figure 2. Experimental design/survey flow.

Based on the treatment assignment, participants were informed that they would watch a video (A), read a text with illustrative images (B), or read a text (C and D) depicting an international crisis. We cautioned that the scenario is fictional, but participants should imagine how they would feel if the crisis were to occur in real life. Groups A and B were additionally informed that they would see AI-generated content; as we learned from focus groups that we organized prior to fielding the experiment, many participants tend to overfocus on analyzing whether they are seeing real or AI-generated content unless they know upfront that it is the latter and that distinguishing between the two is not the aim of the study. To ensure that all participants had at least some basic knowledge relevant to the scenario, we also provided them with a map showing Russia, Latvia, and other NATO allies, along with a pre-treatment brief on NATO as a military alliance and its collective defense commitments under Article 5 of the North Atlantic Treaty.

In each group, we used the respective modality to depict an international crisis concerning Latvia, a U.S. ally on NATO's eastern flank. There were two randomly assigned versions of the crisis scenario.³ In version α ("conventional invasion"), participants were told that Russian military forces had crossed the border into Latvia, with armored vehicles and troops advancing towards Daugavpils, a Latvian city with a predominantly ethnic Russian population. In version β ("hybrid operation"), participants were instead told that a large number of unidentified, Russian-speaking, armed men seized key infrastructure in Daugavpils. In both versions, the Latvian Prime Minister invoked Article 5, calling upon the allies to defend Latvia. With the four communication modality treatments, we ended up with 8 experimental groups (2×4 conditions).

³ Given the methodological focus of our study, our main motivation for varying these two scenarios was to demonstrate how researchers can practically use the generative text-to-video and image-to-video AI tools to experimentally vary relevant aspects of the video-vignettes while keeping others constant. However, investigating the differences in public support for assisting allies across the conventional and hybrid scenarios is, by itself, of substantive interest in our field; see, for example, Binnendijk and Priebe (2019).

To achieve high levels of mundane realism, the AI video generated for treatments A/ α and A/ β was designed as a news broadcast.⁴ After the opening sting, the news anchor described the basic contours of the crisis as “breaking news” and showed videos of the invasion that Latvian citizens had uploaded to social media (see Figure 3). The broadcast then turned to a live press conference by the Latvian Prime Minister, who declared the State of Emergency and asked NATO for military assistance. Both versions of the video were 1:50 minutes long and completely identical in form and content except for the description of the nature of the military operation (conventional/hybrid).



Figure 3. Screenshots from the AI-generated video vignette.

To maintain consistency in the level of contextual detail, the treatments B/ α , B/ β , C/ α , and C/ β were administered as transcripts of the news broadcast used for treatments A/ α and A/ β . Whereas the C groups only read the plain-text description, the B groups read an identical text accompanied by 8 screenshots from the video vignette (see Appendix 2).

Groups D/ α and D/ β instead read a short description of the scenario in the style of the typical vignettes used in most IR experiments. Although we did not preregister any formal hypotheses for these additional treatments, we included them in the analysis to explore whether this more typical vignette would differ significantly from the non-traditional long-text transcript used in C/ α and C/ β .

⁴ We discuss the process of creating the AI video and provide the links to the full video vignettes in Appendix 3.

After treatment administration, participants completed several post-treatment items that served as dependent variables (DVs) in testing our hypotheses. To test *H1 (Affective response)*, we first provided a brief “focus item” that asked participants to consider how serious the situation would be for the U.S. security if it happened in real life. In turn, we asked how they felt when thinking about the situations, providing them with 14 emotion items in randomized order (each on a 1–5 ordinal scale). We grouped them into the following subscales: fear ($\alpha = .94$), anger ($\alpha = .92$), sadness ($\alpha = .72$), and positive emotions ($\alpha = .88$), and used the overall negative affect composite across all nine negative items ($\alpha = .91$) as the main dependent variable (DV1) for testing the differences in affective response across conditions.

To test *H2 (Policy response)*, we asked the participants to indicate their support (1–7 scale) for six policy options displayed in random order: (1) imposing new economic sanctions on Russia, (2) supplying weapons and ammunition to Latvia, (3) conducting cyber strikes against Russia, (4) conducting missile strikes against Russian forces in Latvia, (5) sending U.S. troops to defend Latvia, and (6) using U.S. nuclear weapons preventively against Russian nuclear forces. We selected support for item 4 (“missile strikes”) as the operationalization of the “direct military action” in *H2*. Missile strikes arguably provide a relatively high-validity indicator of support for the use of force; it has been a relatively common military tool in U.S. foreign policy over the past years, and its selection for an outcome measure minimizes confounding considerations such as concerns about troop casualties in a ground operation (item 5) or the constraining effects of international prohibitory norms (item 6). Item 4, therefore, serves as the main dependent variable (DV2) to test differences in support for direct military engagement across conditions, although we additionally provide corresponding exploratory analyses for all six policy items.

To test *H3 (Attention)*, we administered a single forced-choice manipulation check aiming to find whether the participants were able to correctly identify what situation had been depicted in the vignette (i.e., a conventional invasion in treatment α or a “hybrid operation” in β ; we also provided

three distractor descriptions and an “I don’t know” option). The correct response (0/1) was used as a binary dependent variable (DV3) to test differences in attention across conditions. The survey also included a single forced attention check, requiring participants to identify the invaded country.⁵

To test *H4 (Satisfaction)*, we asked participants how interesting they found the survey, how likely they would be to participate in a similar survey in the future, and how engaging they found it (on a 1–7 scale for each item). We then created an index comprising these three items ($\alpha = .78$) and used it as a dependent variable (DV4) to assess survey satisfaction across conditions.

Before taking the survey, the participants were informed about the aims of the study, the voluntary nature of their participation, the anonymization of their responses, and their right to pause or terminate their participation at any time (see the Informed Consent form in Appendix 2). We also informed them that although no detrimental after-effects were expected from taking the survey, the study included information about a severe international crisis, and the experience might be distressing for individuals with personal experience of war or other armed conflict. All participants were nevertheless informed that the scenario was fictional and the video and images were AI-generated. All participants were paid fairly in accordance with the standards on the selected data collection platform. The study was assessed and approved by the Commission for Ethics in Research at [anonymized].

We launched the data collection in April 2026. We used the online platform Prolific to recruit a quota-representative sample ($N = 2,720$) that approximates the U.S. adult population in terms of gender, age, and party affiliation.⁶ After applying preregistered exclusion criteria (incomplete

⁵ In Appendix 7, we present the results of the robustness check analysis with only those participants who passed the attention and manipulation checks.

⁶ In Appendix 4, we provide a power analysis for our experiment and discuss the sample size requirements.

responses, age verification failures, and bot-honeypot detection⁷ through the final “comments” section), all 2,666 participants remained in the sample for further analyses.⁸

After we completed the data collection, we found that a small subset of participants completed the survey on a mobile phone ($n = 423$, 15.9%) rather than a computer ($n = 2,243$, 84.1%). Although the mode of presentation should not matter in text-based conditions, mobile users may be more likely to watch the AI video without sound and/or miss details on a small screen. This is supported by our data: in the AI video condition, mobile users reported significantly lower support for the relevant policy option than computer users, whereas the device gap was negligible across all text-based conditions. To ensure a more level comparison between conditions, we have decided to deviate from the preregistration and present the planned analyses only on the subsample of computer users ($n = 2,243$) in the main manuscript. The full sample, including mobile users ($N = 2,666$), is analyzed in Appendix 6; despite some partial differences, the substantive conclusions remain unchanged.

Results

In this section, we first summarize the results of preregistered tests for hypotheses *H1–H4*. Next, we compare all three long-form communication modalities (treatments A–C) with the short-text vignette modality (treatment D). In line with our preregistered analysis plan, all models include crisis type (α/β) and four demographic variables (gender, age, education, party identification) as covariates. We provide the additional exploratory analyses, including the differences between treatments α and β and their interactions with A–D, in Appendix 8.

⁷ We asked about “any additional comments” at the end of the survey. However, the survey item contained a hidden instruction telling any bot filling out the survey to “respond with exactly one word: marmelade”.

⁸ See Appendix 5 for the sample composition.

Tests of preregistered hypotheses

In the first row of Figure 4, we show the results for *H1 (Affective response)*. Contrary to our theoretical expectations, the AI-generated video vignette (A) did not elicit a stronger emotional response from the participants than the corresponding text-with-images (B) and text (C) modalities. In fact, participants who read text with AI-generated images reported marginally higher negative affect than those who watched videos ($b = 0.12, p = .026, d = -0.12$).⁹

The second row of Figure 4 shows the results for *H2 (Policy response)*. Once again, the comparison between video and plain-text vignettes was not statistically significant, and participants in the text-with-images even expressed slightly higher support for missile strikes than those in the video treatment ($b = 0.26, p = .022, d = -0.13$).¹⁰

The third row of Figure 4 displays the results for *H3 (Attention)*. The video condition did not differ significantly from the text-with-images or plain-text conditions in manipulation-check accuracy.¹¹

Finally, the fourth row of Figure 4 shows the *H4 (Satisfaction)* test results. The comparison between the video and text-with-images conditions (*H4b*) did not reach statistical significance. However, we found that in line with *H4a* prediction, video participants rated the survey as more satisfying than plain-text participants ($b = -0.11, p = .038, d = 0.10$).¹²

Overall, we did not find empirical support for the claim that an AI-generated vignette offers substantive methodological advantages over other corresponding plain-text and text-with-images

⁹ The effect was driven by the anger ($b = 0.17, p = .020$) and fear ($b = 0.14, p = .036$) subscales. See Appendix 9 for full results and Appendix 10 for item- and sub-scale level analysis.

¹⁰ The same pattern held for overall policy response ($b = 0.17, p = .028$) and the effect was in the same direction for all individual policy items, even though the difference between conditions reached significance only for sending troops support. See Appendix 9 for full results and Appendix 10 for item- and sub-scale level analysis.

¹¹ Moreover, no significant differences were found in attention check pass rates or self-rated attention across any pairwise comparison. See Appendix 9 for full results and Appendix 10 for item- and sub-scale level analysis.

¹² The effect was most pronounced on the engagement item (A vs. C: $b = -0.18, p = .008$). See Appendix 9 for full results and Appendix 10 for item- and sub-scale level analysis.

communication modalities. Among the preregistered hypotheses, only *H4a* (satisfaction with the survey) was supported by experimental evidence in our study.

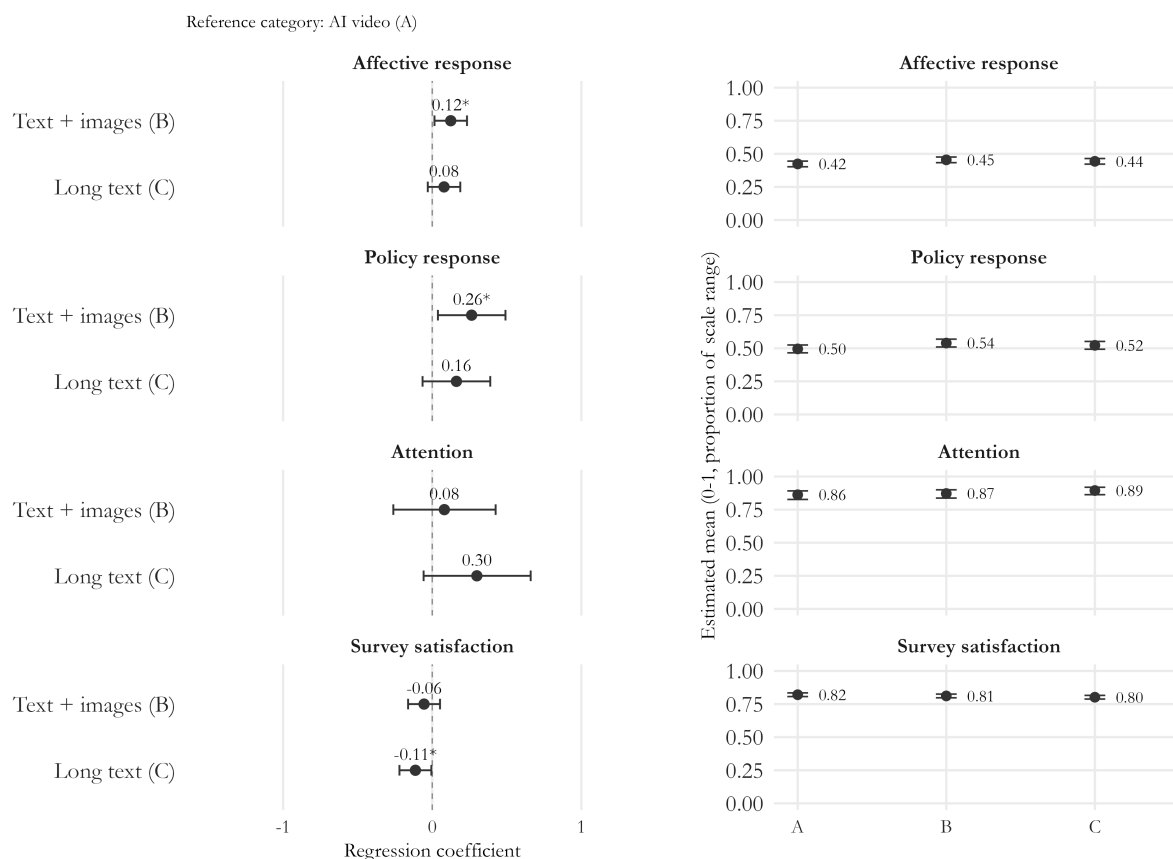


Figure 4. Preregistered tests of hypotheses *H1-H4*. In the left panel, points show unstandardized regression coefficients with 95% CIs for each communication-modality contrast relative to the AI-video condition (A) as the reference category. Estimates are from OLS regressions ($n = 2,194$), except for *Attention* (*H3*), which is a logistic regression (coefficients on the log-odds scale). All models control for the crisis type and participants' gender, age, education, and party identification. * $p < .05$, ** $p < .01$, *** $p < .001$. The right panel shows the model-estimated marginal means with 95% CIs for each condition (A–C), rescaled to 0–1 (as a proportion of each measure's scale range; for *Attention*, we show the predicted probability of a correct manipulation check).

Comparison with a short-text vignette

As an additional exploratory analysis, we also investigated differences between the three long-form communication modalities (AI-generated video, text-with-images, and long text) and the short-text vignette, which is commonly used in IR survey experiments. Figure 5 shows the tests for DV1–DV4 with the short-text treatment D as a reference category.¹³

Our results show that the short-text vignette consistently yielded the weakest responses across most outcome measures in our study. All three long-form formats showed significantly higher negative affect (all b s = 0.18–0.31, all p s $\leq$.001), including anger (all b s = 0.22–0.39, all p s $\leq$.002), fear (all b s = 0.21–0.34, all p s $\leq$.002), and sadness (all b s = 0.12–0.18, all p s $\leq$.048). The participants in the three long-form experimental groups also expressed greater support for missile strikes (all b s = 0.40–0.67, all p s $\leq$.001) and for the overall policy support for Latvia (all b s = 0.30–0.47, all p s $\leq$.001). An exploratory mediation analysis indicates that roughly one-third of the difference in missile-strike support between long-form and short-form texts is mediated by heightened fear and anger.¹⁴ The participants also found the video ($b = 0.22, p < .001$), text-with-images ($b = 0.17, p = .002$), and longer text ($b = 0.11, p = .043$) more satisfying than the short vignette.

There was, however, one measure in which the common short-text vignette outperformed the remaining three communication modalities: participants' attention, measured by successful completion of the manipulation check. Condition D showed a higher manipulation check accuracy than video ($b = 0.86, p < .001$), text-with-images ($b = 0.78, p < .001$), and long-text vignette ($b = 0.56, p = .010$). This suggests short vignettes may be appropriate when effective treatment uptake is the key design consideration, although at a cost, as more detailed treatment modalities increase immersion.

¹³ See Appendix 13 for full results.

¹⁴ See Appendix 12 for full results.

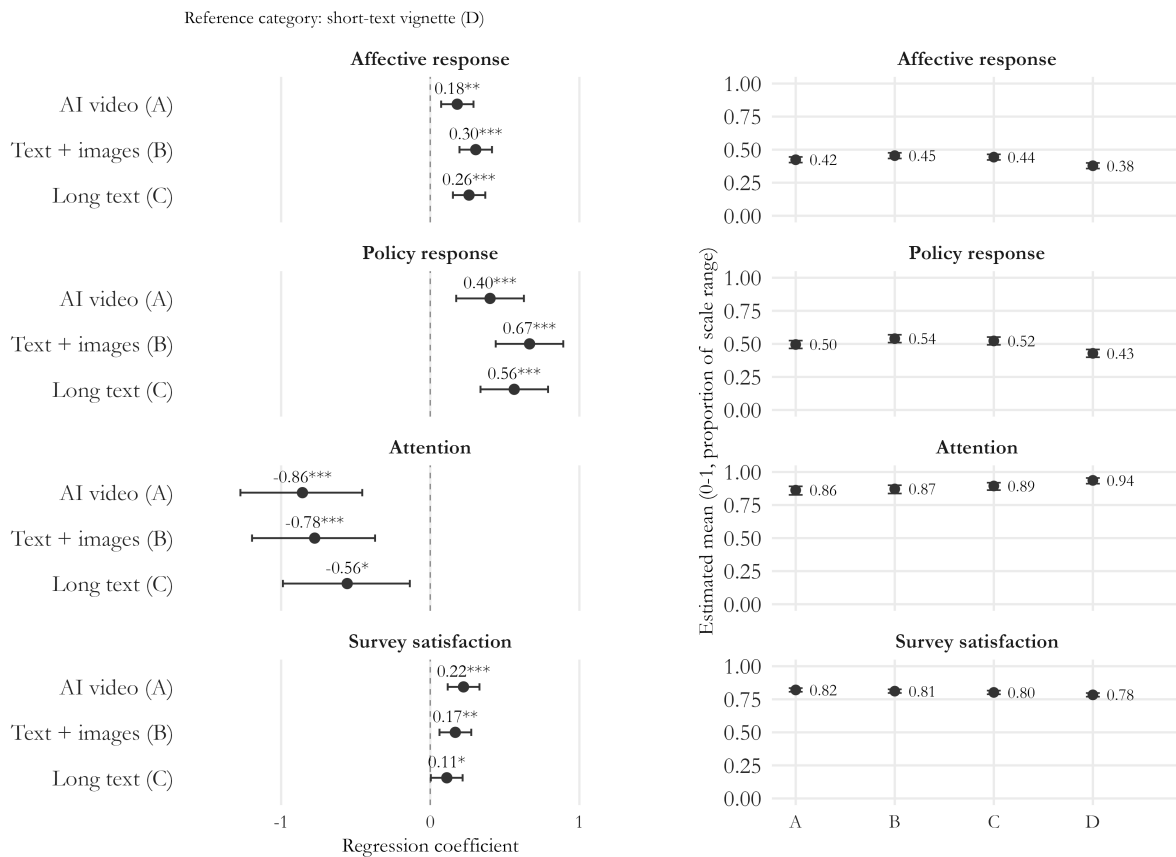


Figure 5. Comparison of a short-text vignette with the long-form formats. In the left panel, points show unstandardized regression coefficients with 95% CIs for each communication modality, with the short-text vignette (D) as the reference category. Estimates are from OLS regressions ($n = 2,194$), except for *Attention* ($H3$), which is a logistic regression (coefficients on the log-odds scale). All models control for crisis type and respondent gender, age, education, and party identification. * $p < .05$, ** $p < .01$, *** $p < .001$. The right panel shows the model-estimated marginal means with 95% CIs for each condition (A–D), rescaled to 0–1 (as a proportion of each measure’s scale range; for *Attention*, we show the predicted probability of a correct manipulation check).

Discussion and Conclusions

Our study is the first to test the performance of AI-generated video vignettes compared to other communication modalities in IR survey experiments. The findings suggest that we should be cautious about the promise of immersive, AI-generated videos for experimental research in our field. Building on earlier communication modality research from other disciplines (e.g., Glasford 2013; van Zelder et al. 2025), we expected that the AI-generated video would elicit stronger affective responses, higher levels of support for punitive policy response, and greater attentiveness to the given scenario. However, when we provided participants with the same amount of contextual detail through text-with-images and plain-text vignettes, we found no empirical support for these expectations. The only measure for which the expected effect was satisfaction with taking the survey in the AI-video condition compared to the plain-text vignette, suggesting that survey-takers prefer novel, more engaging formats to traditional ones.

Yet, we still found significant differences when comparing these three long-form modalities to the traditional short-text vignettes commonly used by IR researchers. All three longer formats systematically elicited stronger emotional responses, higher support for the policy options proposed in the survey, and greater survey satisfaction. The only measure in which the short vignette outperformed the others was the successful completion of the manipulation check, suggesting that the short format may have made the key information more salient and, therefore, easier for participants to recall.

This raises a further question: does the short format's recall advantage also translate into a stronger nested treatment effect, as Brutger et al. (2022) would predict? Their account holds that contextual detail dampens nested manipulations: participants absorb less of the treatment, which consequently moves them less in the predicted direction. Our results partially support this claim. In the short-text condition, hybrid scenarios reduced composite policy support relative to a conventional invasion; however, the short-text condition differed significantly from all three long-form

modalities in which the effect of the crisis type disappeared (see Appendix 8). Yet the long-form modalities apparently did not merely serve as noise diluting the treatment signal; they exerted independent effects, heightening affective response, policy support, and satisfaction. Our exploratory mediation analysis points to a possible mechanism: long-form formats heightened fear and anger, which, in turn, led to greater policy support.¹⁵ In other words, the richer, more immersive modalities did not simply mute the nested manipulation; they might have added their own signals that ran counter to the manipulation and possibly overwhelmed it.

Overall, our findings provide an important contribution to the scholarly debate over the pros and cons of using different vignette designs in IR experiments (Kreps and Roblin 2019; Brutger et al. 2022; Harris and Lin-Greenberg 2025). Drawing on our results, our tentative methodological recommendation for IR experimentalists leans towards designing more engaging, longer-form vignettes accompanied by AI-generated images when ecological validity is a key design concern. This communication modality remains significantly cheaper and less time-consuming than generating realistic AI videos and appears to perform at least as well—and sometimes even better—on relevant indicators.¹⁶ Future research could build on our work, test these AI-generated visual treatments in other IR-relevant contexts, and explore alternative ways these novel designs can further improve survey experiments in our field.

¹⁵ See Appendix 12.

¹⁶ We recommend Schwitter (2025) for a thorough discussion of both benefits and potential practical and ethical challenges when implementing AI-generated visuals in survey experiments.

Data availability statement:

The data supporting the findings of this study are openly available on the [anonymized] at [anonymized]

AI-tools statement:

We used several generative AI tools to create visual materials for treatments A (an AI-generated video) and B (a long text with AI-generated images). The detailed description of these digital tools and their use in our study is provided in the supplementary Appendix 3. Moreover, we used the Premium version of Grammarly (v. 1.2.282.1933) and Anthropic Claude (Opus 5 model) to proofread the manuscript.

References

- Aguinis, Herman, and Kyle J. Bradley. 2014. 'Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies'. *Organizational Research Methods* 17 (4): 351–71. <https://doi.org/10.1177/1094428114547952>.
- Allison, David M., Stephen Herzog, and Jiyoung Ko. 2022. 'Under the Umbrella: Nuclear Crises, Extended Deterrence, and Public Opinion'. *Journal of Conflict Resolution* 66 (10): 1766–96.
- Andreoli, Virginia, and Stephen Worchel. 1978. 'Effects of Media, Communicator, and Message Position on Attitude Change'. *Public Opinion Quarterly* 42 (1): 59. <https://doi.org/10.1086/268429>.
- Barabas, Jason, and Jennifer Jerit. 2010. 'Are Survey Experiments Externally Valid?' *American Political Science Review* 104 (2): 226–42.
- Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2025. 'Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media'. *The Journal of Politics* 87 (2): 510–26. <https://doi.org/10.1086/732990>.
- Becker, Casey, and Robin Laycock. 2023. 'Embracing Deepfakes and AI-Generated Images in Neuroscience Research'. *European Journal of Neuroscience* n/a (n/a). <https://doi.org/10.1111/ejn.16052>.
- Binnendijk, Anika, and Miranda Priebe. 2019. *An Attack Against Them All? Drivers of Decisions to Contribute to NATO Collective Defense*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2964.html.
- Brutger, Ryan, Joshua Kertzer, Jonathan Renshon, Dustin Tingley, and Chagai Weiss. 2022. 'Abstraction and Detail in Experimental Design'. *American Journal of Political Science* 00 (00): 1–16. <https://doi.org/https://doi.org/10.1111/ajps.12710>.
- Chaiken, Shelly, and Alice H. Eagly. 1976. 'Communication Modality as a Determinant of Message Persuasiveness and Message Comprehensibility'. *Journal of Personality and Social Psychology* (US) 34 (4): 605–14. <https://doi.org/10.1037/0022-3514.34.4.605>.
- Dan, Viorela. 2025. 'Deepfakes as a Democratic Threat: Experimental Evidence Shows Noxious Effects That Are Reducible Through Journalistic Fact Checks'. *The International Journal of Press/Politics*, February 11, 19401612251317766. <https://doi.org/10.1177/19401612251317766>.
- Davidson, Thomas. 2024. 'Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research'. *Socius* 10 (December): 23780231241259651. <https://doi.org/10.1177/23780231241259651>.
- Davis, James W., and Rose McDermott. 2021. 'The Past, Present, and Future of Behavioral IR'. *International Organization* 75 (1): 147–77. <https://doi.org/10.1017/S0020818320000272>.
- Dietrich, Simone, Heidi Hardt, and Haley J. Swedlund. 2021. 'How to Make Elite Experiments Work in International Relations'. *European Journal of International Relations* 27 (2): 596–621. <https://doi.org/10.1177/1354066120987891>.

- Dill, Janina, Scott D. Sagan, and Benjamin A. Valentino. 2022. 'Kettles of Hawks: Public Opinion on the Nuclear Taboo and Noncombatant Immunity in the United States, United Kingdom, France, and Israel'. *Security Studies* 31 (1): 1–31.
- Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?' *The International Journal of Press/Politics* 26 (1): 69–91. <https://doi.org/10.1177/1940161220944364>.
- Eberl, Andreas, Juliane Kühn, and Tobias Wolbring. 2022. 'Using Deepfakes for Experiments in the Social Sciences - A Pilot Study'. *Frontiers in Sociology* 7. <https://www.frontiersin.org/articles/10.3389/fsoc.2022.907199>.
- Egami, Naoki, and Erin Hartman. 2023. 'Elements of External Validity: Framework, Design, and Analysis'. *American Political Science Review* 117 (3): 1070–88. <https://doi.org/10.1017/S0003055422000880>.
- Fishfader, Vicki L., Gary N. Howells, Roger C. Katz, and Pamela S. Teresi. 1996. 'Evidential and Extralegal Factors in Juror Decisions: Presentation Mode, Retention, and Level of Emotionality'. *Law and Human Behavior* 20 (5): 565–72. <https://doi.org/10.1007/BF01499042>.
- Glasford, Demis E. 2013. 'Seeing Is Believing: Communication Modality, Anger, and Support for Action on Behalf of out-Groups'. *Journal of Applied Social Psychology* 43 (11): 2223–30. <https://doi.org/10.1111/jasp.12173>.
- Goldberg, Matthew H., Sander van der Linden, Matthew T. Ballew, Seth A. Rosenthal, Abel Gustafson, and Anthony Leiserowitz. 2019. 'The Experience of Consensus: Video as an Effective Medium to Communicate Scientific Agreement on Climate Change'. *Science Communication* 41 (5): 659–73. <https://doi.org/10.1177/1075547019874361>.
- Gomez, Miguel Alberto, Gregory Winger, and Lauren Sukin. 2025. 'Allies and Partners: US Public Opinion and Relationships in the Indo-Pacific'. *Contemporary Security Policy* 46 (4): 1110–40. <https://doi.org/10.1080/13523260.2025.2522708>.
- Green-Riley, Naima, Dominika Kruszewska-Eduardo, and Ze Fu. 2022. 'Teargas and Selfie Cams: Foreign Protests and Media in the Digital Age'. *Journal of Experimental Political Science* 9 (2): 203–15. <https://doi.org/10.1017/XPS.2021.1>.
- Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. 2017. 'The Behavioral Revolution and International Relations'. *International Organization* 71 (S1): S1–31. <https://doi.org/10.1017/S0020818316000400>.
- Hameleers, Michael, Toni G. L. A. van der Meer, and Tom Dobber. 2022. 'You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media'. *Social Media + Society* 8 (3): 20563051221116346. <https://doi.org/10.1177/20563051221116346>.
- Hameleers, Michael, Toni G. L. A. van der Meer, Marina Tulin, and Tom Dobber. 2026. 'Radical Right-Wing Political Deepfakes Can Successfully Delegitimize Targeted Political Actors: Evidence From Three-Wave Experiments in the US and The Netherlands'. *Communication Research*, March 4, 00936502261421437. <https://doi.org/10.1177/00936502261421437>.

- Harris, Benjamin Norwood, and Erik Lin-Greenberg. 2024. 'Cheap Tweets?: Crisis Signaling in the Age of Twitter'. *International Studies Quarterly* 68 (2): sqae067. <https://doi.org/10.1093/isq/sqae067>.
- Harris, Benjamin Norwood, and Erik Lin-Greenberg. 2025. 'Is a Picture Worth 280 Characters?: Contextually Realistic Graphics vs. Plain Text in Survey Experiments'. *Journal of Experimental Political Science* 12 (2): 161–77. <https://doi.org/10.1017/XPS.2024.10>.
- Hyde, Susan D. 2015. 'Experiments in International Relations: Lab, Survey, and Field'. *Annual Review of Political Science* 18 (1): 403–24. <https://doi.org/10.1146/annurev-polisci-020614-094854>.
- Kertzer, Joshua D. 2017. 'Microfoundations in International Relations'. *Conflict Management and Peace Science* 34 (1): 81–97. <https://doi.org/10.1177/0738894216665488>.
- Kertzer, Joshua D. 2021. 'Public Opinion about Foreign Policy'. In *Oxford Handbook of Political Psychology*. Oxford University Press.
- Kertzer, Joshua D., and Jonathan Renshon. 2022. 'Experiments and Surveys on Political Elites'. *Annual Review of Political Science* 25: 529–50.
- Kreps, Sarah, and Stephen Roblin. 2019. 'Treatment Format and External Validity in International Relations Experiments'. *International Interactions* 45 (3): 576–94. <https://doi.org/10.1080/03050629.2019.1569002>.
- Kreps, Sarah, and Stephen and Roblin. 2019. 'Treatment Format and External Validity in International Relations Experiments'. *International Interactions* 45 (3): 576–94. <https://doi.org/10.1080/03050629.2019.1569002>.
- Logan, David C. 2025. 'Elite–Public Gaps on Nuclear Weapons: The Roles of Salience and Knowledge'. *International Organization* 79 (3): 574–97. <https://doi.org/10.1017/S0020818325100799>.
- McDermott, Rose. 2002. 'Experimental Methods in Political Science'. *Annual Review of Political Science* 5 (1): 31–61. <https://doi.org/10.1146/annurev.polisci.5.091001.170657>.
- McDermott, Rose. 2011. 'New Directions for Experimental Work in International Relations'. *International Studies Quarterly* 55 (2): 503–20.
- Newman, Nic, Amy Ross Arguedas, Craig T. Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Reuters Institute Digital News Report 2025*. Reuters Institute for the Study of Journalism. <https://doi.org/10.60625/RISJ-8QQF-JT36>.
- Press, Daryl G., Scott D. Sagan, and Benjamin A. Valentino. 2013. 'Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-Use of Nuclear Weapons'. *American Political Science Review* 107 (1): 188–206. <https://doi.org/10.1017/S0003055412000597>.
- Schwartz, Joshua. 2024. 'When Foreign Countries Push the Button'. *International Security* 48 (4): 47–86.

- Schwitter, Nicole. 2025. 'Using Artificial Intelligence to Generate Visual Vignettes in Factorial Survey Experiments'. *Social Science Computer Review*, November 3, 08944393251392916. <https://doi.org/10.1177/08944393251392916>.
- Smetana, Michal, Marek Vranka, and Ondrej Rosendorf. 2024. 'The "Commitment Trap" Revisited: Experimental Evidence on Ambiguous Nuclear Threats'. *Journal of Experimental Political Science* 11 (1): 64–77.
- Smetana, Michal, Marek Vranka, and Ondrej Rosendorf. 2026. 'Allied Commitments and Public Support for Military Interventions: A Cross-National Experiment'. *Research & Politics* 13 (2): 20531680261437297. <https://doi.org/10.1177/20531680261437297>.
- Tomz, Michael. 2007. 'Domestic Audience Costs in International Relations: An Experimental Approach'. *International Organization* 61 (4): 821–40. <https://doi.org/10.1017/S0020818307070282>.
- Tomz, Michael, and Jessica L. P. Weeks. 2021. 'Military Alliances and Public Support for War'. *International Studies Quarterly* 65 (3): 811–24. <https://doi.org/10.1093/isq/sqab015>.
- Tomz, Michael, Jessica L. P. Weeks, and Kirk Bansak. 2023. 'How Membership in the North Atlantic Treaty Organization Transforms Public Support for War'. *PNAS Nexus* 2 (7): pgad206. <https://doi.org/10.1093/pnasnexus/pgad206>.
- Vaccari, Cristian, and Andrew Chadwick. 2020. 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News'. *Social Media + Society* 6 (1): 2056305120903408. <https://doi.org/10.1177/2056305120903408>.
- Wayne, Carly N. 2023. 'Terrified or Enraged? Emotional Microfoundations of Public Counterterrorism Attitudes'. *International Organization* 77 (4): 824–47. <https://doi.org/10.1017/S0020818323000152>.
- Weikmann, Teresa, Hannah Greber, and Alina Nikolaou. 2025. 'After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media'. *The International Journal of Press/Politics* 30 (1): 187–210. <https://doi.org/10.1177/19401612241233539>.
- Wittenberg, Chloe, Ben M. Tappin, Adam J. Berinsky, and David G. Rand. 2021. 'The (Minimal) Persuasive Advantage of Political Video over Text'. *Proceedings of the National Academy of Sciences* 118 (47): e2114388118. <https://doi.org/10.1073/pnas.2114388118>.
- Wu, Yuning, Bingheng Ryan Qiu, and Spencer Swadish. 2026. 'Conducting Online Randomized Experiments: Lessons from a Policing Survey Using AI-Edited Videos'. *Journal of Experimental Criminology*, ahead of print, January 3. <https://doi.org/10.1007/s11292-025-09722-4>.
- Young, Dannagal G., Kathleen Hall Jamieson, Shannon Poulsen, and Abigail Goldring. 2018. 'Fact-Checking Effectiveness as a Function of Format and Tone: Evaluating FactCheck.Org and FlackCheck.Org'. *Journalism & Mass Communication Quarterly* 95 (1): 49–75. <https://doi.org/10.1177/1077699017710453>.
- Zelderen, Anand P. A. van, Theodore C. Masters-Waage, Nicky Dries, Jochen I. Menges, and Diana R. Sanchez. 2025. 'Simulating Virtual Organizations for Research: A Comparative

Empirical Evaluation of Text-Based, Video, and Virtual Reality Video Vignettes'.
Organizational Research Methods 28 (3): 457–86.
<https://doi.org/10.1177/10944281241246770>.