Supplementary Materials

Delegating Destruction: Coercive Threats and Automated Nuclear Systems

Joshua A. Schwartz* & Michael C. Horowitz†

Contents

Robustness Tests **1**
Study 1 . . . 1
Study 2 . . . 3

Heterogeneous Effects **6**
Study 1 . . . 6
Study 2 . . . 9
Study 3 . . . 11

Explicit Nuclear Threats vs. Implicit Nuclear Threats **13**
Background . . . 13
Study 1 . . . 18
Study 2 . . . 21
Study 3 . . . 23

Questionnaire (Studies 1 and 2) **24**

Questionnaire (Study 3) **36**

Pre-Registered Hypotheses **39**

References **43**

*Assistant Professor of International Relations and Emerging Technology, Carnegie Mellon Institute for Strategy and Technology, Carnegie Mellon University, joshschwartz@cmu.edu.

†Richard Perry Professor of Political Science and Director of Perry World House, Department of Political Science, University of Pennsylvania, horom@sas.upenn.edu.

Robustness Tests

Study 1

Table A.1 shows that increased threat credibility and effectiveness in the automated treatment compared to the non-automated treatment is robust to several tests. First, in models 1 and 2, we use the full-scale version of our dependent variables. Second, in models 3 and 4, we utilize binary versions of our dependent variables. The coefficient in model 4 is slightly over conventional levels of statistical significance when utilizing a two-sided test ($p = 0.131$), but the statistically significant result when using the full scale of this dependent variable, which we contend is our best measure since it enables us to exploit the maximum amount of variation, builds confidence that the difference between the automated and non-automated treatment is real. Third, in models 5 and 6, we drop respondents that failed a pre-treatment attention check. Finally, in models 7 and 8, we control for a range of covariates. Many of the covariates also have effects that are intuitive. For instance, model 8 shows that greater hawkishness, support for NATO Article 5 commitments, and support for military aid to Ukraine is associated with a statistically significant decrease in threat effectiveness; that is, willingness to give in to Putin's demands.

Table A.1: Study 1 Robustness Tests

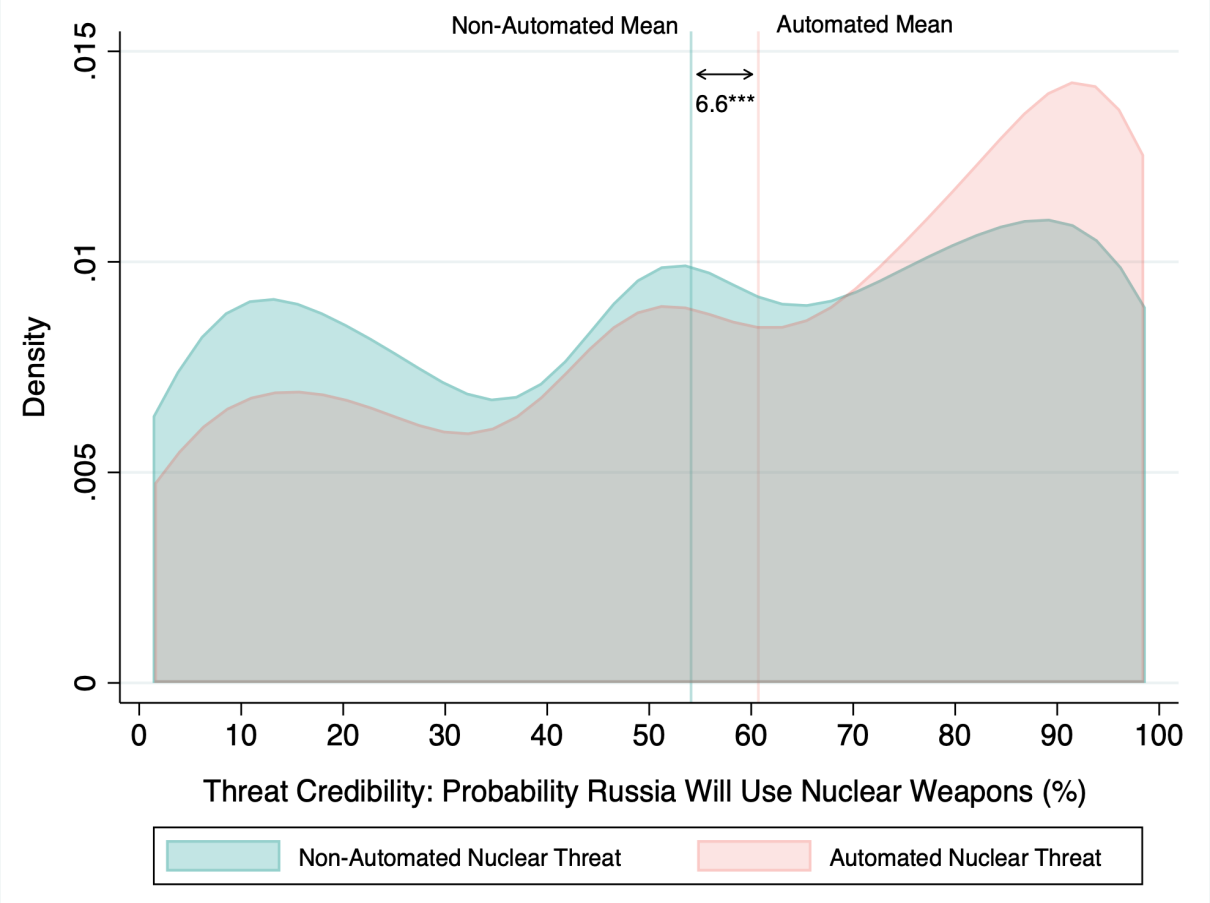
	(1) Cred	(2) Effect	(3) Cred	(4) Effect	(5) Cred	(6) Effect	(7) Cred	(8) Effect
Automated Versus Non-Automated Threat	12.3430***	0.2809**	0.1486***	0.0524	12.8795***	0.2878**	12.2773***	0.2480**
	(2.2903)	(0.1408)	(0.0348)	(0.0347)	(2.3454)	(0.1443)	(2.1746)	(0.1190)
Labour Party							-0.5494	-0.1512
							(2.4782)	(0.1366)
Conservative Ideology							-0.6371	-0.1270**
							(1.0983)	(0.0576)
Hawkishness							0.5078	-0.3790***
							(1.6211)	(0.0929)
Support for NATO Article 5							0.4048	-0.2921***
							(1.5622)	(0.0875)
Support for Ukraine Military Aid							0.7872	-0.4191***
							(1.2200)	(0.0725)
AI Knowledge							-1.0232	-0.1688**
							(1.4996)	(0.0843)
Attention to World Affairs							4.6075***	0.0400
							(1.3825)	(0.0764)
Education							-1.7232	-0.0691
							(1.1867)	(0.0583)
Age							-0.4930***	-0.0235***
							(0.0783)	(0.0044)
Income							0.2378	0.0066
							(0.4874)	(0.0263)
Region							1.6435	0.0964
							(1.7348)	(0.0925)
Female							9.1808***	0.4999***
							(2.4010)	(0.1367)
White							-2.9705	0.0673
							(3.5594)	(0.1997)
Constant	51.8076***	3.8228***	0.4835***	0.3772***	51.4933***	3.8356***	65.4381***	9.6367***
	(1.6320)	(0.0952)	(0.0252)	(0.0244)	(1.6745)	(0.0982)	(12.8344)	(0.6364)
Observations	800	800	800	800	760	760	800	800
Full Dependent Variable Scale	✓	✓	×	×	✓	✓	✓	✓
Binary Dependent Variable Scale	×	×	✓	✓	×	×	×	×
Excludes Inattentive Respondents	×	×	×	×	✓	✓	×	×

Notes: Robust standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

Study 2

Figure A.1 plots the densities for the automated and non-automated nuclear threat treatments and shows visually that the perceived probability that Russia will use nuclear weapons is over 6.5 percentage points greater in the automated than the non-automated nuclear threat treatment in Study 2. The effect size is smaller, as expected due to the weaker treatments, in Study 2 (6.6 percentage points) than in Study 1 (12.3 percentage points), but remains statistically significant at the 0.01 level. The robustness of this finding even to weaker treatments builds confidence in our results related to threat credibility.

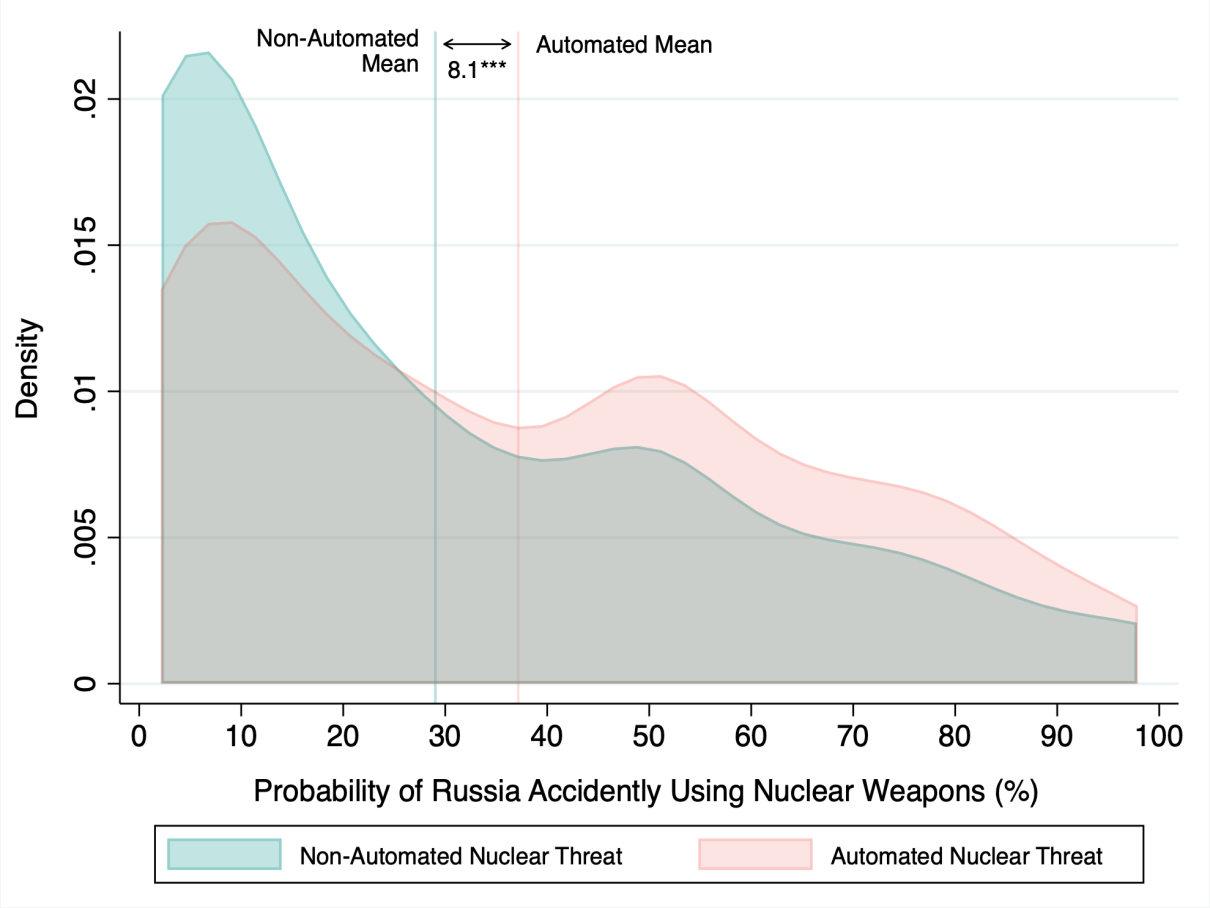
Figure A.1: Automated Nuclear Threats Enhance Credibility (Study 2)



Note: * = p<0.10, ** = p<0.05, and *** = p<0.01.

Figure A.2 demonstrates that our finding related to the perceived chance of a nuclear accident is robust to using weaker treatments in Study 2. Respondents continue to believe that automated nuclear threats are more likely to lead to the mistaken use of nuclear weapons than non-automated nuclear threats.

Figure A.2: Automated Nuclear Threats Increase Perceived Risks of Accidents (Study 2)



Note: * = p<0.10, ** = p<0.05, and *** = p<0.01.

Table A.2 shows that increased threat credibility in the automated treatment compared to the non-automated treatment is robust to several tests. First, in model 1, we use the full-scale version of our dependent variables. Second, in model 3, we utilize binary versions of our dependent variables. Third, in model 5, we drop respondents that failed a pre-treatment attention check. Finally, in model 7, we control for a range of covariates. The difference in threat effectiveness between the automated and non-automated treatments remains statistically insignificant for these tests, likely due to the weaker nature of the treatment. Many of the covariates also have effects that are intuitive. For instance, model 8 shows that greater hawkishness, support for NATO Article 5 commitments, and support for military aid to Ukraine is associated with a statistically significant decrease in threat effectiveness; that is, willingness to give in to Putin's demands.

Table A.2: Study 2 Robustness Tests

	(1) Cred	(2) Effect	(3) Cred	(4) Effect	(5) Cred	(6) Effect	(7) Cred	(8) Effect
Automated Versus Non-Automated Threat	6.6020***	0.0679	0.0759**	0.0295	6.1103***	0.0715	6.9711***	0.0834
	(1.9363)	(0.1166)	(0.0304)	(0.0303)	(1.9734)	(0.1190)	(1.8211)	(0.1022)
Labour Party							1.1103	-0.1208
							(2.0924)	(0.1161)
Conservative Ideology							-0.9303	-0.0903*
							(0.9259)	(0.0518)
Hawkishness							2.4522*	-0.4514***
							(1.2921)	(0.0689)
Support for NATO Article 5							2.9613**	-0.1892**
							(1.3011)	(0.0741)
Support for Ukraine Military Aid							-2.4758**	-0.3554***
							(1.1341)	(0.0632)
AI Knowledge							-1.3472	-0.1635**
							(1.1722)	(0.0652)
Attention to World Affairs							4.5803***	0.0881
							(1.0858)	(0.0608)
Education							-1.4625	-0.0487
							(1.0183)	(0.0531)
Age							-0.4406***	-0.0067*
							(0.0702)	(0.0040)
Income							0.1216	-0.0105
							(0.3803)	(0.0217)
Region							1.8173	-0.0254
							(1.3077)	(0.0859)
Female							9.9889***	0.3924***
							(2.0128)	(0.1103)
White							-2.9815	-0.2375
							(2.8586)	(0.1588)
Constant	54.1022***	4.0373***	0.5285***	0.3988***	54.3245***	4.0306***	61.6395***	8.7757***
	(1.4015)	(0.0837)	(0.0221)	(0.0217)	(1.4223)	(0.0855)	(10.2134)	(0.5764)
Observations	1060	1060	1060	1060	1019	1019	1060	1060
Full Dependent Variable Scale	✓	✓	×	×	✓	✓	✓	✓
Binary Dependent Variable Scale	×	×	✓	✓	×	×	×	×
Excludes Inattentive Respondents	×	×	×	×	✓	✓	×	×

Notes: Robust standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

Heterogeneous Effects

Study 1

In our pre-analysis plan, we hypothesized that the difference in threat credibility and effectiveness between the automated and non-automated treatments would be larger for respondents that pay closer attention to world affairs. The logic being that these respondents may be more familiar with the fact that Putin has made nuclear threats in the past and not followed through on them, which could make the non-automated threat in particular (which corresponds more closely to Putin's real past threats) less credible and effective. We test this expectation in Table A.3. Our pre-registered hypothesis would be supported if the coefficients on the interaction effects between the automated threat treatment and our attention to world affairs variables are positive and statistically significant. We find only mixed support for this expectation. Our hypothesis is supported only in model 2, which analyzes threat credibility using a binary measure of attention to world affairs.[1]

[1]This variable is coded a 1 if respondents pay attention to international affairs "most of the time."

Table A.3: Study 1 Heterogeneous Effects—Attention to World Affairs

	(1) Cred	(2) Cred	(3) Effect	(4) Effect
Automated Threat × Attention to World Affairs	-3.0946		0.0318	
	(2.4417)		(0.1315)	
Automated Threat × Attention to World Affairs (Binary)		11.7740**		0.1935
		(5.0251)		(0.2676)
Attention to World Affairs	6.1849***		0.0238	
	(1.8786)		(0.0986)	
Attention to World Affairs (Binary)		-16.8341***		-0.2865
		(3.5759)		(0.1916)
Automated Threat	19.0299***	9.3019***	0.1785	0.1995
	(5.7150)	(2.4911)	(0.3044)	(0.1389)
Labour Party	-0.4349	-0.3314	-0.1524	-0.1477
	(2.4775)	(2.4693)	(0.1370)	(0.1366)
Conservative Ideology	-0.5634	-0.3785	-0.1278**	-0.1262**
	(1.0944)	(1.0890)	(0.0576)	(0.0572)
Hawkishness	0.4463	0.4243	-0.3783***	-0.3780***
	(1.6059)	(1.5975)	(0.0931)	(0.0924)
Support for NATO Article 5	0.4892	0.3360	-0.2929***	-0.2881***
	(1.5581)	(1.5635)	(0.0875)	(0.0874)
Support for Ukraine Military Aid	0.7194	0.8116	-0.4184***	-0.4174***
	(1.2165)	(1.2006)	(0.0725)	(0.0725)
AI Knowledge	-0.9614	-1.3553	-0.1694**	-0.1653**
	(1.4927)	(1.4529)	(0.0844)	(0.0827)
Education	-1.7089	-1.6480	-0.0693	-0.0637
	(1.1867)	(1.1816)	(0.0584)	(0.0583)
Age	-0.4920***	-0.5057***	-0.0235***	-0.0230***
	(0.0780)	(0.0766)	(0.0044)	(0.0042)
Income	0.2565	0.2656	0.0064	0.0078
	(0.4860)	(0.4838)	(0.0263)	(0.0263)
Region	1.6233	1.2948	0.0966	0.0926
	(1.7434)	(1.7413)	(0.0926)	(0.0928)
Female	9.1669***	9.0227***	0.5001***	0.4851***
	(2.3999)	(2.4104)	(0.1368)	(0.1361)
White	-2.8276	-2.8143	0.0658	0.0701
	(3.5649)	(3.5281)	(0.2000)	(0.1996)
Constant	61.3791***	80.3842***	9.6784***	9.7105***
	(13.0250)	(11.3558)	(0.6523)	(0.5560)
Observations	800	800	800	800
Full Dependent Variable Scale	✓	✓	✓	✓

Notes: Robust standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

In Table A.4 we test other potential heterogeneous effects that may impact the difference in threat credibility or effectiveness between the automated and non-automated treatments. We test whether political identification (models 1 and 2), hawkishness (models 3 and 4), support for NATO's Article 5 (models 5 and 6), support for military aid to Ukraine (models 7 and 8), and self-reported AI knowledge (models 9 and 10) act as moderators. The interaction coefficients reveal no statistically significant effects. The only exception is model 4, where the effectiveness of automated threats relative to non-automated threats is smaller for more hawkish respondents. This is likely due to the fact that more hawkish respondents are dispositionally more likely to be opposed to giving in to coercive threats, which makes even automated nuclear threats less likely to impact their views. Nevertheless, as discussed below, this heterogeneous effect does not hold in Study 2.

Table A.4: Study 1 Heterogeneous Effects—Party Identification, Hawkishness, Support for NATO and Aid to Ukraine, and AI Knowledge

	(1) Cred	(2) Effect	(3) Cred	(4) Effect	(5) Cred	(6) Effect	(7) Cred	(8) Effect	(9) Cred	(10) Effect
Automated Threat × Labour Party	2.8606	0.2147								
	(4.3527)	(0.2428)								
Automated Threat × Hawkishness			0.0330	-0.3660***						
			(2.6297)	(0.1367)						
Automated Threat × Support for NATO Article 5					1.7691	-0.0247				
					(2.6755)	(0.1396)				
Automated Threat × Support for Ukraine Military Aid							-0.2103	-0.0711		
							(2.0628)	(0.1117)		
Automated Threat × AI Knowledge									0.8771	-0.1126
									(2.5945)	(0.1451)
Automated Threat	11.1250***	0.1615	12.1744	1.3911***	4.8193	0.3520	13.1094	0.5294	10.3381*	0.4968
	(2.8875)	(0.1506)	(8.4279)	(0.4425)	(11.5307)	(0.5965)	(8.5294)	(0.4566)	(6.0162)	(0.3323)
Labour Party	-1.9811	-0.2586	-0.5489	-0.1575	-0.6010	-0.1505	-0.5507	-0.1516	-0.5807	-0.1472
	(3.2376)	(0.1749)	(2.4809)	(0.1364)	(2.4803)	(0.1367)	(2.4796)	(0.1368)	(2.4762)	(0.1366)
Conservative Ideology	-0.6538	-0.1283**	-0.6372	-0.1260**	-0.6135	-0.1273**	-0.6420	-0.1287**	-0.6425	-0.1263**
	(1.0977)	(0.0576)	(1.0993)	(0.0578)	(1.0986)	(0.0577)	(1.0998)	(0.0579)	(1.0984)	(0.0576)
Hawkishness	0.5425	-0.3764***	0.4927	-0.2107*	0.4811	-0.3786***	0.5135	-0.3771***	0.5201	-0.3805***
	(1.6265)	(0.0934)	(1.9416)	(0.1134)	(1.6195)	(0.0931)	(1.6204)	(0.0930)	(1.6219)	(0.0926)
Support for NATO Article 5	0.3635	-0.2952***	0.4039	-0.2819***	-0.5600	-0.2786**	0.4032	-0.2926***	0.4111	-0.2929***
	(1.5622)	(0.0877)	(1.5674)	(0.0863)	(2.1646)	(0.1153)	(1.5620)	(0.0874)	(1.5646)	(0.0876)
Support for Ukraine Military Aid	0.7715	-0.4203***	0.7869	-0.4148***	0.8134	-0.4194***	0.8961	-0.3822***	0.7916	-0.4196***
	(1.2213)	(0.0726)	(1.2201)	(0.0725)	(1.2234)	(0.0728)	(1.6448)	(0.0880)	(1.2196)	(0.0726)
AI Knowledge	-1.0329	-0.1695**	-1.0227	-0.1741**	-0.9950	-0.1692**	-1.0295	-0.1709**	-1.4816	-0.1100
	(1.5020)	(0.0844)	(1.5006)	(0.0840)	(1.4988)	(0.0847)	(1.5029)	(0.0848)	(1.9276)	(0.1139)
Attention to World Affairs	4.5826***	0.0382	4.6082***	0.0329	4.5685***	0.0406	4.6016***	0.0380	4.6066***	0.0402
	(1.3821)	(0.0764)	(1.3856)	(0.0759)	(1.3806)	(0.0765)	(1.3869)	(0.0765)	(1.3832)	(0.0764)
Education	-1.7308	-0.0697	-1.7229	-0.0724	-1.7337	-0.0690	-1.7233	-0.0692	-1.7024	-0.0718
	(1.1886)	(0.0584)	(1.1879)	(0.0579)	(1.1909)	(0.0584)	(1.1871)	(0.0583)	(1.1943)	(0.0587)
Age	-0.4932***	-0.0235***	-0.4930***	-0.0239***	-0.4933***	-0.0235***	-0.4936***	-0.0237***	-0.4919***	-0.0236***
	(0.0783)	(0.0044)	(0.0782)	(0.0043)	(0.0782)	(0.0044)	(0.0786)	(0.0044)	(0.0784)	(0.0044)
Income	0.2331	0.0062	0.2381	0.0039	0.2516	0.0064	0.2365	0.0061	0.2429	0.0059
	(0.4887)	(0.0263)	(0.4876)	(0.0261)	(0.4873)	(0.0263)	(0.4886)	(0.0263)	(0.4873)	(0.0262)
Region	1.6432	0.0963	1.6423	0.1094	1.5816	0.0972	1.6539	0.0999	1.6497	0.0956
	(1.7292)	(0.0920)	(1.7381)	(0.0937)	(1.7484)	(0.0932)	(1.7432)	(0.0940)	(1.7387)	(0.0927)
Female	9.2392***	0.5043***	9.1792***	0.5172***	9.2065***	0.4996***	9.1859***	0.5017***	9.1478***	0.5042***
	(2.4032)	(0.1367)	(2.4075)	(0.1360)	(2.4028)	(0.1369)	(2.4035)	(0.1365)	(2.4027)	(0.1373)
White	-2.8050	0.0797	-2.9711	0.0739	-3.1302	0.0695	-2.9447	0.0760	-2.9954	0.0705
	(3.5677)	(0.1989)	(3.5618)	(0.1990)	(3.5977)	(0.2010)	(3.5655)	(0.2004)	(3.5594)	(0.1998)
Constant	66.1972***	9.6937***	65.4867***	9.0968***	69.6775***	9.5776***	65.0348***	9.5003***	66.2656***	9.5305***
	(12.8763)	(0.6418)	(13.3995)	(0.6616)	(14.4750)	(0.7039)	(13.3629)	(0.6575)	(12.8626)	(0.6372)
Observations	800	800	800	800	800	800	800	800	800	800
Full Dependent Variable Scale	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Robust standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

Study 2

As discussed previously, in our pre-analysis plan we hypothesized that the difference in threat credibility and effectiveness between the automated and non-automated treatments would be larger for respondents that pay closer attention to world affairs. We test this expectation in Table A.5. Our pre-registered hypothesis would be supported if the coefficients on the interaction effects between the automated threat treatment and our attention to world affairs variables are positive and statistically significant. We find no statistically significant evidence for this expectation.

Table A.5: Study 2 Heterogeneous Effects—Attention to World Affairs

	(1) Cred	(2) Cred	(3) Effect	(4) Effect
Automated Threat × Attention to World Affairs	-2.9972		0.0226	
	(1.9191)		(0.1056)	
Automated Threat × Attention to World Affairs (Binary)		6.2985		0.2259
		(4.1698)		(0.2224)
Attention to World Affairs	6.0790***		0.0768	
	(1.4107)		(0.0821)	
Attention to World Affairs (Binary)		-10.7972***		-0.3498**
		(3.1756)		(0.1669)
Automated Threat	13.3480***	5.1491**	0.0353	0.0195
	(4.5844)	(2.1200)	(0.2476)	(0.1210)
Labour Party	1.3410	1.2069	-0.1225	-0.1191
	(2.0971)	(2.1020)	(0.1161)	(0.1160)
Conservative Ideology	-0.9026	-0.8352	-0.0905*	-0.0908*
	(0.9266)	(0.9280)	(0.0518)	(0.0518)
Hawkishness	2.3178*	2.2793*	-0.4504***	-0.4560***
	(1.2987)	(1.2959)	(0.0691)	(0.0688)
Support for NATO Article 5	2.9693**	2.6047**	-0.1893**	-0.1954***
	(1.3052)	(1.3154)	(0.0741)	(0.0736)
Support for Ukraine Military Aid	-2.4440**	-2.4649**	-0.3557***	-0.3508***
	(1.1373)	(1.1414)	(0.0633)	(0.0632)
AI Knowledge	-1.2886	-1.6034	-0.1639**	-0.1605**
	(1.1746)	(1.1737)	(0.0652)	(0.0654)
Education	-1.4732	-1.6675*	-0.0486	-0.0496
	(1.0215)	(1.0123)	(0.0531)	(0.0526)
Age	-0.4419***	-0.4695***	-0.0066*	-0.0064
	(0.0701)	(0.0690)	(0.0040)	(0.0039)
Income	0.0998	0.0977	-0.0103	-0.0100
	(0.3799)	(0.3833)	(0.0218)	(0.0218)
Region	1.7819	1.8217	-0.0252	-0.0240
	(1.3020)	(1.3154)	(0.0859)	(0.0856)
Female	9.9636***	10.4772***	0.3926***	0.3931***
	(2.0139)	(2.0116)	(0.1103)	(0.1101)
White	-3.0054	-2.6925	-0.2373	-0.2328
	(2.8432)	(2.8469)	(0.1588)	(0.1593)
Constant	58.6777***	78.4734***	8.7980***	9.0597***
	(10.3002)	(9.0364)	(0.5856)	(0.5093)
Observations	1060	1060	1060	1060
Full Dependent Variable Scale	✓	✓	✓	✓

Notes: Robust standard errors in parentheses. *p<0.10; **p<0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

In Table A.6 we test other potential heterogeneous effects that may impact the difference in threat credibility or effectiveness between the automated and non-automated treatment in Study 2. We test whether political identification (models 1 and 2), hawkishness (models 3 and 4), support for NATO's Article 5 (models 5 and 6), support for military aid to Ukraine (models 7 and 8), and self-reported AI knowledge (models 9 and 10) act as moderators. The interaction coefficients reveal no statistically significant effects for Study 2.

Table A.6: Study 2 Heterogeneous Effects—Party Identification, Hawkishness, Support for NATO and Aid to Ukraine, and AI Knowledge

	(1) Cred	(2) Effect	(3) Cred	(4) Effect	(5) Cred	(6) Effect	(7) Cred	(8) Effect	(9) Cred	(10) Effect
Automated Threat × Labour Party	-1.5480	0.0188								
	(3.7560)	(0.2099)								
Automated Threat × Hawkishness			1.4233	0.0361						
			(2.1137)	(0.1105)						
Automated Threat × Support for NATO Article 5					2.0419	-0.0549				
					(1.9989)	(0.1075)				
Automated Threat × Support for Ukraine Military Aid							1.7153	0.0538		
							(1.7481)	(0.0925)		
Automated Threat × AI Knowledge									2.4110	-0.1457
									(2.0915)	(0.1178)
Automated Threat	7.5777***	0.0760	2.4789	-0.0305	-1.2729	0.3051	0.4342	-0.1216	1.2461	0.4293
	(2.3808)	(0.1293)	(6.8882)	(0.3583)	(8.3232)	(0.4405)	(6.9047)	(0.3652)	(5.2492)	(0.2945)
Labour Party	1.9023	-0.1304	1.0636	-0.1220	1.1421	-0.1216	1.0623	-0.1223	1.1074	-0.1206
	(2.8509)	(0.1553)	(2.0944)	(0.1163)	(2.0930)	(0.1160)	(2.0946)	(0.1161)	(2.0912)	(0.1160)
Conservative Ideology	-0.9303	-0.0903*	-0.9380	-0.0905*	-0.9350	-0.0902*	-0.9118	-0.0897*	-0.8528	-0.0950*
	(0.9257)	(0.0518)	(0.9255)	(0.0518)	(0.9282)	(0.0518)	(0.9263)	(0.0518)	(0.9332)	(0.0518)
Hawkishness	2.4653*	-0.4515***	1.7353	-0.4696***	2.3934*	-0.4498***	2.3640*	-0.4542***	2.3866*	-0.4474***
	(1.2930)	(0.0690)	(1.6211)	(0.0831)	(1.2898)	(0.0691)	(1.2948)	(0.0689)	(1.2967)	(0.0688)
Support for NATO Article 5	2.9338**	-0.1889**	2.9504**	-0.1895**	1.8979	-0.1606*	2.9843**	-0.1885**	3.0645**	-0.1955***
	(1.3030)	(0.0741)	(1.3041)	(0.0741)	(1.6432)	(0.0856)	(1.3062)	(0.0739)	(1.3034)	(0.0743)
Support for Ukraine Military Aid	-2.4479**	-0.3558***	-2.4892**	-0.3558***	-2.4492**	-0.3561***	-3.3579**	-0.3831***	-2.4710**	-0.3557***
	(1.1394)	(0.0632)	(1.1338)	(0.0633)	(1.1326)	(0.0631)	(1.4447)	(0.0752)	(1.1323)	(0.0632)
AI Knowledge	-1.3280	-0.1637**	-1.3369	-0.1632**	-1.2322	-0.1666**	-1.2946	-0.1619**	-2.5852*	-0.0887
	(1.1747)	(0.0651)	(1.1715)	(0.0652)	(1.1782)	(0.0657)	(1.1706)	(0.0652)	(1.5557)	(0.0870)
Attention to World Affairs	4.6181***	0.0876	4.6317***	0.0894	4.6218***	0.0870	4.5972***	0.0886	4.5326***	0.0910
	(1.0962)	(0.0608)	(1.0860)	(0.0611)	(1.0880)	(0.0610)	(1.0888)	(0.0608)	(1.0872)	(0.0609)
Education	-1.4685	-0.0486	-1.4592	-0.0486	-1.4277	-0.0497	-1.4505	-0.0483	-1.4514	-0.0494
	(1.0181)	(0.0531)	(1.0192)	(0.0532)	(1.0214)	(0.0530)	(1.0204)	(0.0531)	(1.0169)	(0.0531)
Age	-0.4409***	-0.0067*	-0.4396***	-0.0066*	-0.4365***	-0.0068*	-0.4386***	-0.0066*	-0.4417***	-0.0066*
	(0.0701)	(0.0040)	(0.0701)	(0.0040)	(0.0702)	(0.0040)	(0.0702)	(0.0040)	(0.0702)	(0.0040)
Income	0.1196	-0.0104	0.1125	-0.0107	0.0904	-0.0096	0.1069	-0.0109	0.0999	-0.0092
	(0.3804)	(0.0217)	(0.3821)	(0.0217)	(0.3809)	(0.0217)	(0.3806)	(0.0217)	(0.3802)	(0.0218)
Region	1.8589	-0.0260	1.7924	-0.0261	1.7226	-0.0229	1.8312	-0.0250	1.7986	-0.0243
	(1.3105)	(0.0860)	(1.3039)	(0.0860)	(1.3126)	(0.0861)	(1.3068)	(0.0857)	(1.3070)	(0.0861)
Female	9.9675***	0.3926***	9.9665***	0.3918***	9.9798***	0.3926***	9.9365***	0.3907***	10.0585***	0.3882***
	(2.0138)	(0.1102)	(2.0135)	(0.1103)	(2.0142)	(0.1103)	(2.0143)	(0.1105)	(2.0197)	(0.1098)
White	-2.9844	-0.2375	-3.0595	-0.2395	-2.8664	-0.2406	-2.8592	-0.2337	-3.1067	-0.2299
	(2.8622)	(0.1588)	(2.8628)	(0.1586)	(2.8572)	(0.1595)	(2.8540)	(0.1591)	(2.8835)	(0.1591)
Constant	61.1620***	8.7815***	63.9913***	8.8354***	65.5133***	8.6716***	64.8320***	8.8758***	64.3746***	8.6105***
	(10.3327)	(0.5807)	(10.6961)	(0.5946)	(11.0509)	(0.6098)	(10.7054)	(0.6116)	(10.4473)	(0.5917)
Observations	1060	1060	1060	1060	1060	1060	1060	1060	1060	1060
Full Dependent Variable Scale	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Robust standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01. Cred = Credibility, Effect = Effectiveness.

Study 3

Table A.7 tests whether there are any heterogeneous effects for our core results in Study 3. In other words, are there factors that increase or decrease the difference in threat credibility, threat effectiveness, and the perceived likelihood of accidents between the automated and non-automated treatments? Note that since this is a fully within-subject experiment, there is no treatment to interact potential moderators with. Instead, evidence for heterogeneous effects can be found in a simple regression where the dependent variable is the within-subject effect and a series of covariates are included.[2] A significant effect for any of the covariates would indicate that the within-subject effect is stronger for different values of that covariate. We find little consistent evidence for heterogeneous effects.

Table A.7: Study 3 Heterogeneous Effects

	(1) Credibility	(2) Effectiveness	(3) Accident
Labour Party	-7.955	0.048	2.751
	(6.880)	(0.228)	(2.750)
Conservative Party	-10.554	-0.125	6.522
	(7.489)	(0.231)	(4.300)
Woman MP	-5.467	-0.182	-10.613*
	(4.579)	(0.270)	(5.789)
Region	-0.224	-0.037	0.266
	(0.972)	(0.050)	(0.683)
Year Elected	-0.242	0.020	1.100
	(1.237)	(0.091)	(1.538)
Non-Automated Treatment First	-5.037	0.250	-1.915
	(4.171)	(0.225)	(3.791)
Constant	12.870	0.150	-2.476
	(8.271)	(0.377)	(6.818)
Observations	84	89	76

Notes: Standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01.

[2] YouGov only provides a small number of covariates for UK MP studies to maintain the respondents' anonymity.

A non-trivial percentage of UK MPs (about 23% on average) "preferred not to answer" some of our outcome measure questions, which was an option YouGov insisted we give respondents. Some respondents might have answered this way because they objected to responding to any questions about a hypothetical foreign policy situation in general. Others might have objected to *specific* questions because they felt answering honestly would be embarrassing or have other negative effects. We assess this latter possibility in Table A.8 by analyzing whether "preferred not to answer" responses were more common in the automated versus the non-automated treatment. We find no statistical evidence that it was. This provides suggestive evidence that the main reason for respondents indicating they "preferred not to answer" was a more generalized opposition to answering our questions rather than being impacted by the treatments themselves.

Table A.8: Prefer Not to Answer Responses

	(1) Credibility	(2) Effectiveness
Difference Between Automated and Non-Automated Treatment	0.000 (0.024)	-0.009 (0.015)
Observations	117	117

Notes: Standard errors in parentheses. *p<0.10; **p< 0.05; ***p<0.01.

Explicit Nuclear Threats vs. Implicit Nuclear Threats

Background

Although the main focus of our article is whether nuclear threats implemented via automation are more credible and effective than those implemented without, we also tested the efficacy of explicit nuclear threats compared to implicit threats. This ties into two debates in the literature. First, as discussed in the main text, are nuclear threats credible and effective at all outside of deterring major aggression against one's homeland? Second, what is the efficacy of explicit threats relative to more implicit ones? As defined by the Cambridge Dictionary, something explicit is "communicated directly in a clear and exact way," whereas something implicit is "suggested but not communicated directly." Jankovic and Roeben (2025, 9) further explain the difference between explicit and implicit threats:

> "Explicit threats often convey the gravity of the potential consequences in a clear and direct manner. The threatening state might specify the extent of the force it intends to use, such as deploying nuclear weapons or launching a full-scale invasion. In the case of implicit threats, the gravity is less clear and must be inferred from actions, context, or the threatening state's capabilities, positioning of forces, or past conduct. Also, in explicit threats the timing is often clearly communicated, whereas implicit threats often lack a clear timeline, creating uncertainty about when or if the threatened action will occur."

Some scholars and practitioners hint that implicit nuclear threats may be just as effective as explicit ones. Kissinger (1956, 351), for example, said, "overt threats have become unnecessary; every calculation of risks will have to include the Soviet stockpile of atomic weapons and ballistic missiles." On the other hand, Snyder and Diesing (1977, 216) contend that "maximum explicitness and clarity in threats tends to produce maximum credibility", even if leaders have historically been hesitant to make this kind of explicit threat for fear that it will reduce their future flexibility. We agree with this latter argument. In fact, we contend comparing explicit to implicit nuclear threats is a relatively easy test for finding evidence of the coercive efficacy of explicit nuclear threats. Implicit nuclear threats, which do not make the gravity of non-compliance clear and force

the target to infer the danger of nuclear use from a country's capabilities, are weaker than another type of threat—ambiguous nuclear threats. This type of coercion does not explicitly threaten nuclear use, but makes clearer that if the target does not comply, then they will face grave consequences (Sagan, 2000). Logically, one of those consequences could be nuclear use. In other words, on the spectrum from most explicit to no threat, implicit threats are even less explicit than ambiguous threats.

Our expectation that explicit nuclear threats will be more credible and effective than implicit threats draws from audience cost theory, which holds that public threats can signal resolve because if leaders fail to follow through on their threats and act inconsistently, then they will face punishment from domestic audiences (Fearon, 1997). Previous work even demonstrates that audience costs may operate in both democracies and autocracies (Weeks, 2008; Li and Chen, 2021; Smetana, 2025). Although autocratic leaders do not need to worry about public opinion as much as democratic ones—perhaps rendering audience costs among the general public in autocracies less relevant—a growing body of literature demonstrates that autocratic regimes do respond to public opinion due to the fear of rebellion and other politics costs (Chen, Pan, and Xu, 2016; Truex, 2016; Lueders, 2022). They may also face audience costs among elite actors (e.g., other members of the regime) for backing down from threats. Given that not upholding an explicit threat may be more likely to be perceived as backing down than not following through on an implicit threat, we expected that explicit nuclear threats—whether implemented with or without the aid of automated nuclear systems—should be more credible in the eyes of a country's adversary than implicit threats. The logic being that (a) if a leader is more likely to be punished by domestic audiences for not using nuclear weapons after issuing an explicit nuclear threat, then that leader should be more likely to use nuclear weapons to avoid domestic punishment; and (b) only a relatively resolved leader would be willing to tie their hands in this way in the first place and risk domestic punishment. Indeed, Smetana, Vranka, and Rosendorf (2024) find empirical evidence that leaders are punished more when they back down from explicit threats than ambiguous threats. Given that implicit threats are even less clear than ambiguous ones, we thus expect the difference

between explicit and implicit threats could be even stronger. A finding in accordance with our pre-registered expectations would show that explicit threats are more effective than implicit ones, and provide evidence that nuclear coercion can be effective outside of deterring major aggression against one's homeland.

While we hypothesized that explicit nuclear threats provide some advantages in coercive bargaining, we expected that they are not without drawbacks. In particular, by raising the risks of nuclear use, explicit threats are likely to increase threat perceptions in target audiences due to the fear of escalation. This is especially the case given the existence of a norm against the use of nuclear weapons—even if its strength is highly debatable (Dill, Sagan, and Valentino, 2022; Schwartz, 2024)—which will likely make an explicit nuclear threat for an offensive purpose particularly shocking and threatening in the eyes of target audiences (Tannenwald, 1999; Paul, 2009). Heightened threat perceptions may then impact the kinds of policy preferences held by target audiences. In particular, it might incentivize balancing behavior—such as increasing military spending and maintaining a nuclear deterrent—to address the threat.

To test our hypotheses related to explicit versus implicit nuclear threats, we designed a hybrid within and between-subjects experiment modeled off of a study by Yarhi-Milo, Kertzer, and Renshon (2018). There are three key experimental conditions: (1) a baseline condition where an implicit public threat is made, (2) a treatment condition where an explicit non-automated nuclear threat is made, and (3) a treatment condition where an explicit automated nuclear threat is made. In the baseline condition that all respondents receive, Vladimir Putin makes an implicit public threat that does not explicitly threaten nuclear use, though Russia's nuclear capabilities remain hovering in the background (Kissinger, 1956):

> "Vladimir Putin has also issued a public statement warning NATO not to launch any missile attacks against Russian forces in Estonia, or deploy any strike aircraft in operations against Russian forces. If NATO does either of these things, then Russia's early-warning radar and satellite systems could detect it."

Note that the threat here is entirely implicit, as Putin does not actually threaten to *do* anything. Although he warns NATO not to take an action, the gravity of vio-

lating this red line is totally unclear and must simply be inferred from the context of the scenario and Russia's capabilities (e.g., its nuclear arsenal, which respondents are reminded about before receiving the implicit threat). This implicit threat differs from and is weaker than ambiguous nuclear threats that have been tested in prior literature. For example, Smetana, Vranka, and Rosendorf (2024) operationalize an ambiguous nuclear threat by having the president say the following: "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!" Compared to our implicit threat, this ambiguous threat is more specific about the gravity of the consequences if the red line is violated—"overwhelming and devastating"—even if it is not fully explicit that nuclear weapons will be used. Consequently, as discussed above, this is an easier test for finding a significant effect for explicit threats compared to if Putin made an ambiguous nuclear threat.

Following the Yarhi-Milo, Kertzer, and Renshon (2018) study, we then inform respondents they will be given an alternative version of the scenario they just read, tell them the basic details will remain the same, and randomly assign them to one of two treatments—(1) the explicit automated nuclear threat treatment and (2) the explicit non-automated nuclear threat treatment. Respondents are then asked a series of dependent variable questions *both* after the baseline scenario *and again* after whichever treatment they are assigned to. The within-subject effect between the explicit threat treatments and the baseline condition where an implicit threat is made is the quantity of interest for this analysis.

Within-subject experiments are valid tools for causal inference in this case, despite theoretical concerns about demand effects and consistency pressures. Demand effects occur if respondents surmise researchers' hypotheses and adjust their behavior to validate those expectations. Within-subject designs could potentially increase the chance of demand effects because the repeated measure may alert respondents to what researchers are testing. However, a comprehensive study conducted by Mummolo and Peterson (2019) found minimal evidence for the existence of demand effects. They experimentally ma-

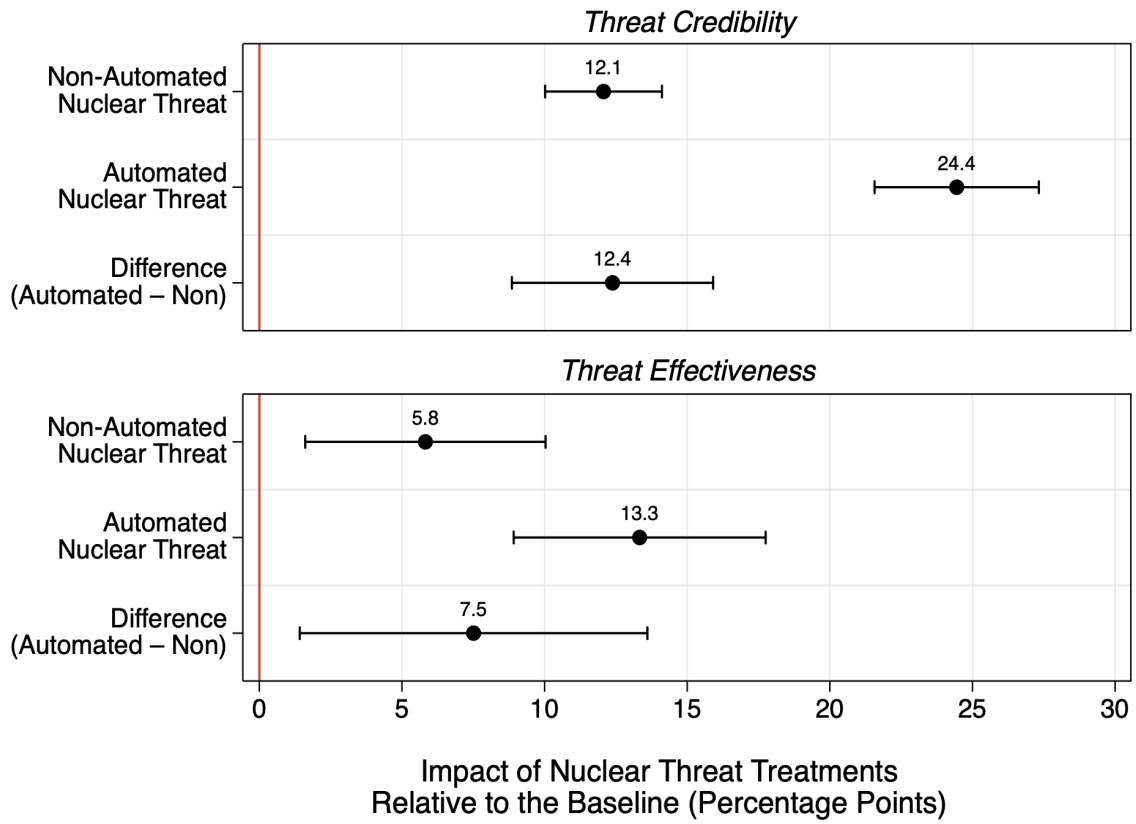
nipulated the amount of information provided to survey subjects about the researchers' goals and hypotheses, and they found it had little impact on treatment effects. Another possibility is that within-subject designs could lead to consistency pressures, where respondents feel the need to respond to dependent variable questions later in an experiment in a similar way as they responded to them earlier in the experiment. If anything, this would bias our study against finding any effects, making it a harder test. Nevertheless, Clifford, Sheagley, and Piston (2021) demonstrate that this concern—in addition to worries about demand effects—are overstated. They directly compare between-subject and within-subject designs and find they yield substantively similar results, but that within-subject designs substantially increase statistical power and precision. The gains in statistical power can be "dramatic," yielding standard errors that are between 20% and 58% lower than comparable between-subject designs (Clifford, Sheagley, and Piston, 2021). Consequently, on net, the evidence suggests that within-subject designs, like the one used here, provide valid and credible estimates of treatment effects.

Study 1

Per our theoretical expectations, Figure A.3 illustrates that in Study 1 both automated and non-automated explicit nuclear threats significantly increase the perceived probability that Russia will use nuclear weapons if their red line is violated compared to an implicit nuclear threat. For example, while only 35% of respondents believe that Russia would likely use a nuclear weapon (i.e., the probability of use is over 50%) if their red line was violated following an implicit threat, a majority of respondents (56%) believe Russia would likely use a nuclear weapon following the violation of an explicit nuclear threat. Furthermore, our two explicit nuclear threat treatments also increase the effectiveness of Putin's threat compared to the baseline implicit threat condition. That is, they make the UK public less willing to violate Putin's red line by employing missiles or strike aircraft to defend Estonia. For example, while a majority of respondents (54%) are willing to support violating Putin's red line to use missiles and strike aircraft to defend Estonia following an implicit threat by Putin, less than a majority (45%) are willing to do so following an explicit nuclear threat. The impact of explicit threats is so strong that it reduces willingness to engage in *any kind* of military action to protect Estonia, even military action that does not violate the letter of Putin's threat by utilizing missiles or strike aircraft. This may be because they are worried about an accidental trigger of the automated system. These findings demonstrate that explicit nuclear threats pass a relatively easy test of their credibility. It further shows that nuclear threats do have some ability to aid in offensive and revisionist efforts, such as a Russian invasion of Estonia. These results are robust to utilizing a binary measure of credibility, the full 7-point measure of threat effectiveness, or dropping respondents that failed a pre-treatment attention check.[3]

[3]Control variables are not needed for this analysis given that it is a within-subject test and respondents serve as their own control.

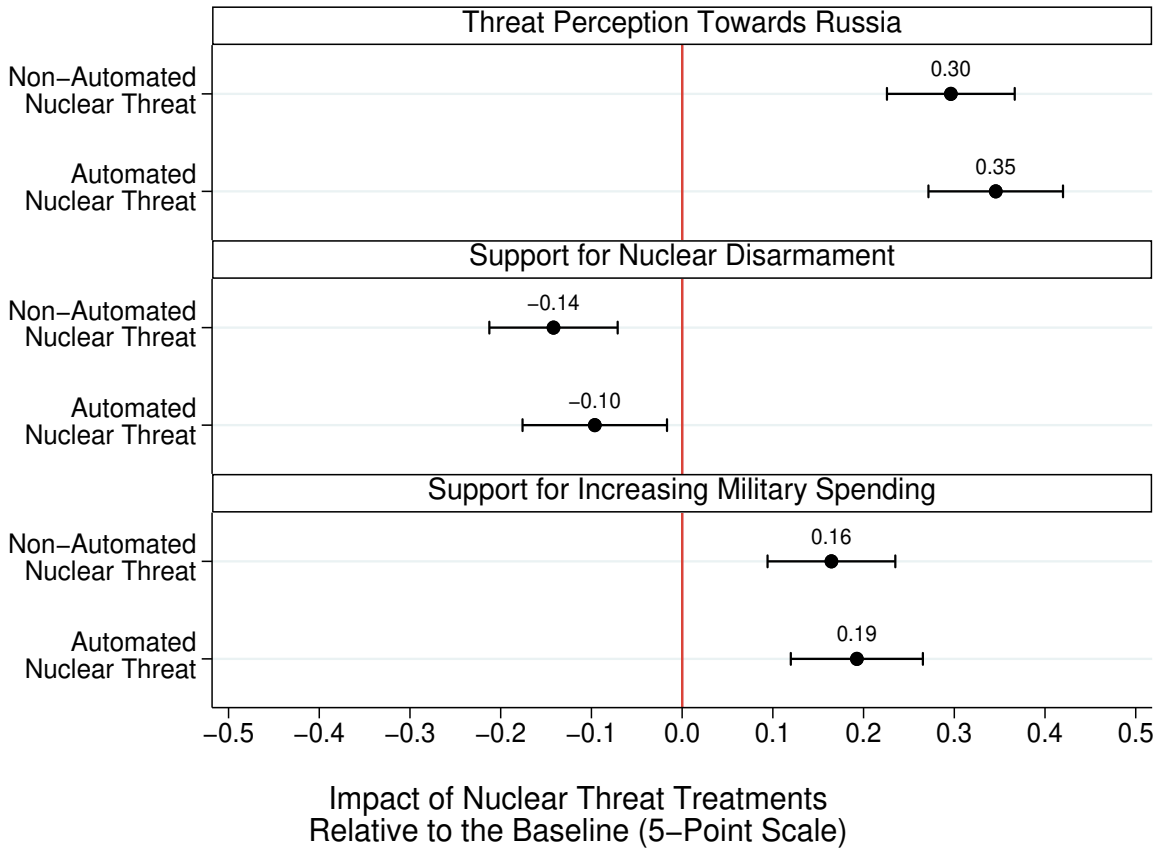
Figure A.3: Advantages of Explicit Threats Relative to Implicit Threats



Note: Bars are 95% confidence intervals.

Nevertheless, we also find that there are downsides associated with making explicit nuclear threats relative to implicit threats. As shown in Figure A.4, explicit nuclear threats (a) increase threat perception towards Russia and (b) UK public willingness to "significantly" increase defense spending relative to an implicit threat. Explicit threats also reduce support for nuclear disarmament relative to implicit threats. Therefore, issuing explicit nuclear threats is not a panacea or costless for the coercer.

Figure A.4: Disadvantages of Explicit Threats Relative to Implicit Threats

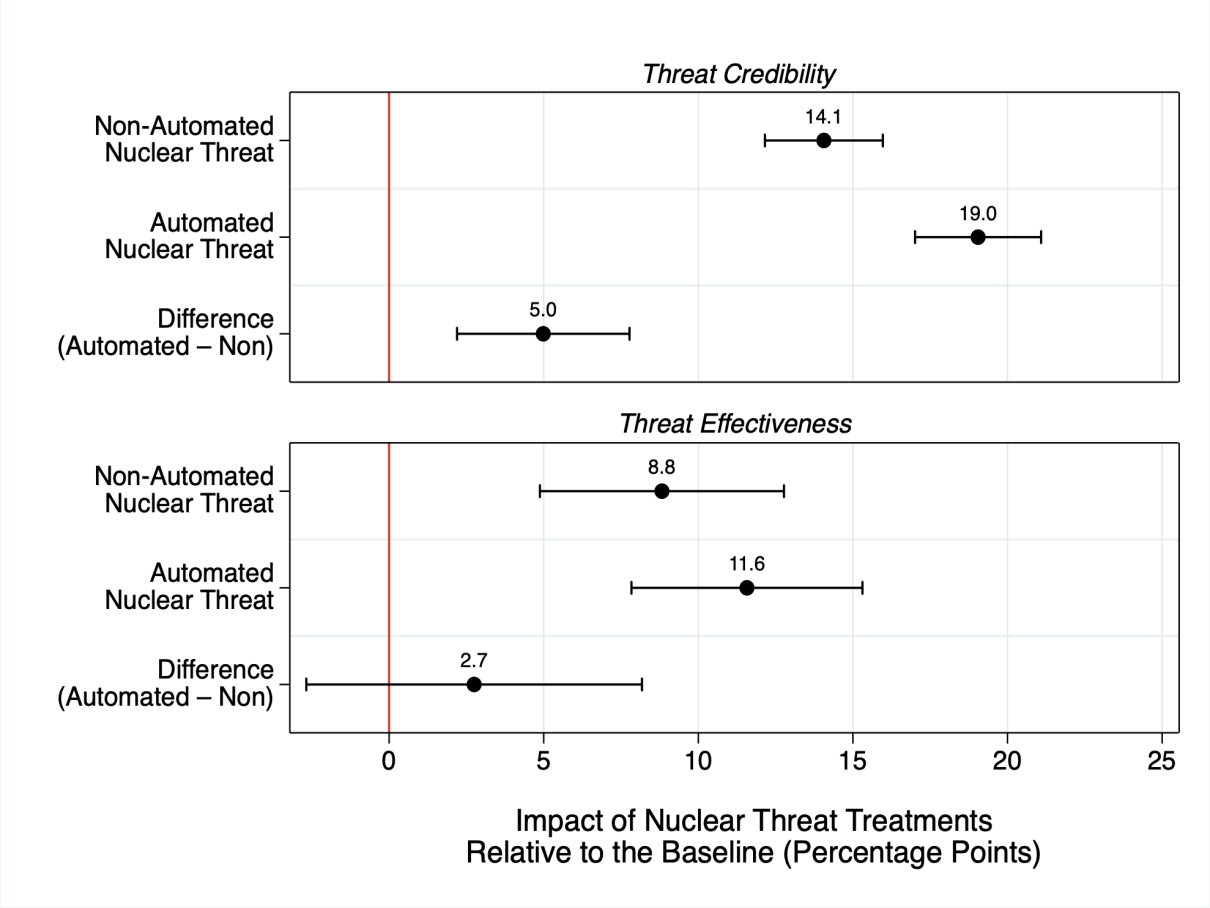


Note: Bars are 95% confidence intervals.

Study 2

Per the results in Study 1, threat credibility and threat effectiveness in Study 2 are greater in the explicit nuclear threat conditions than in the baseline condition where an implicit threat is made.

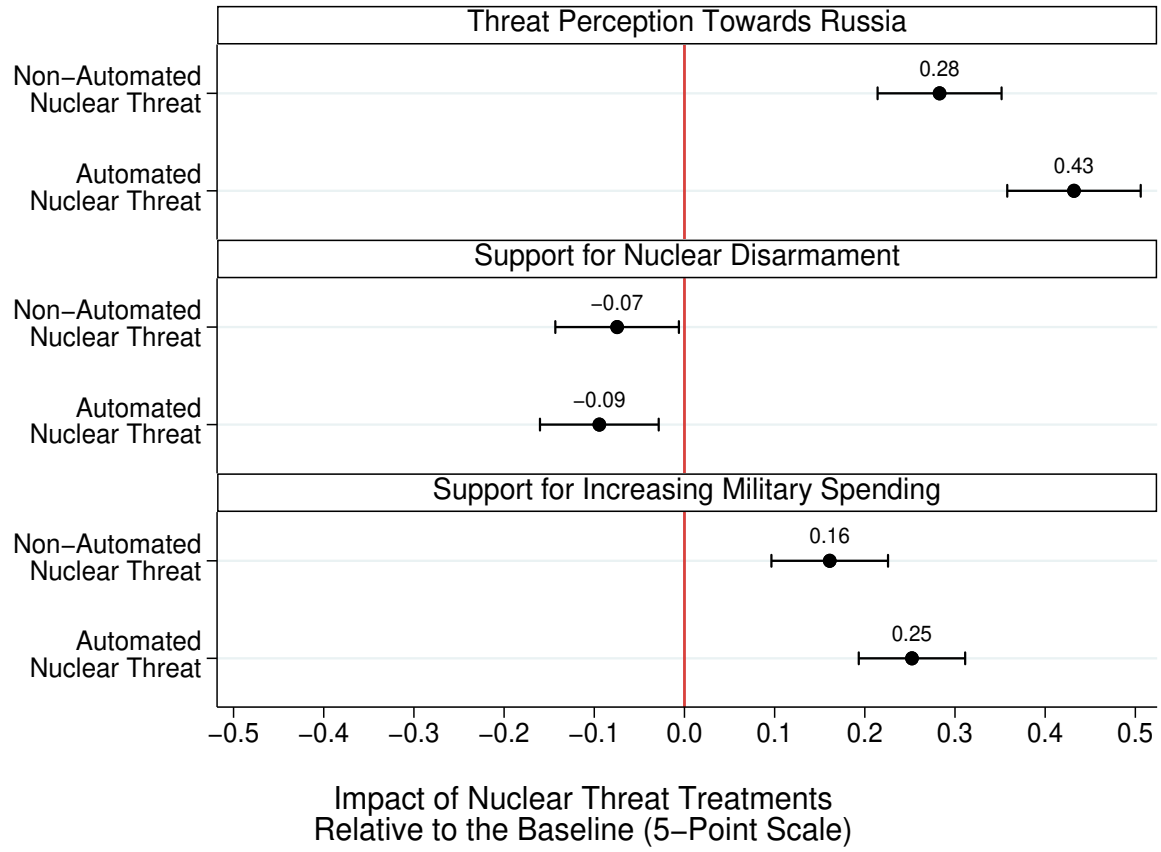
Figure A.5: Advantages of Explicit Threats Relative to Implicit Threats



Note: Bars are 95% confidence intervals.

As in Study 1, threat perceptions towards Russia and support for increasing military spending is greater when an explicit nuclear threat is made than when an implicit threat is issued, and support for nuclear disarmament is lower.

Figure A.6: Disadvantages of Explicit Threats Relative to Implicit Threats



Note: Bars are 95% confidence intervals.

Study 3

In accordance with our results among the UK public, we find that the credibility and effectiveness of Russian nuclear threats among UK MPs is significantly higher when Putin makes an explicit nuclear threat—whether automated or non-automated—compared to an implicit nuclear threat.[4] However, although the effects are statistically significant, their substantive size is quite moderate when using the full scale of our dependent variable measures, especially relative to the effect sizes in the public studies. Alternatively, some of the substantive effects are much larger when operationalizing the dependent variable in a different way. For example, if we collapse the 7-point measure of threat effectiveness into a binary variable, the percentage of UK MPs who would support violating Putin's red line is 13.3 percentage points ($p < 0.01$) less when Putin makes an explicit nuclear threat implemented via automation than when he makes an implicit threat. This is a large effect, especially given that the within-subject design of the experiment may have put consistency pressure on UK MPs, incentivizing them to respond similarly to questions about an automated nuclear threat from Russia and an implicit threat from Russia. In sum, we find strong and consistent evidence among both elites and the public that explicit nuclear threats are more credible and effective than implicit threats.

Table A.9: UK Members of Parliament Results

	Implicit Nuclear Threat	Explicit Non-Automated Nuclear Threat	Explicit Automated Nuclear Threat	Difference Non-Automated vs. Implicit	Difference Automated vs. Implicit
Threat Credibility	34.3%	38.5%	39.1%	3.62pp**	4.45pp**
Threat Effectiveness	2.5	2.8	3.0	0.39***	0.56***

Note: * = p<0.10, ** = p<0.05, and *** = p<0.01, where p-values indicate whether differences are statistically greater than 0. PP = Percentage Points. The results for threat effectiveness are on a 7-point scale.

[4] We find no consistent evidence of heterogeneous effects.

Questionnaire (Studies 1 and 2)

Informed Consent[5]

Do you consent to participate in this study?

Participation in this study is completely voluntary. You are free to decline to participate, or to end participation at any time for any reason.

This study will present you with a hypothetical scenario and ask your opinions about it. It should take about 10 minutes. There are no anticipated risks associated with your participation. Although this study will not benefit you personally, we hope that our results will add to knowledge about attitudes toward international relations.

The study is also anonymous, meaning we will not collect information that could be used to identify you, such as your name or personal address. The data we collect may be made available to other researchers online at some point, but your responses will remain anonymous.

I have read the consent form, and I want to continue with this study	I do not want to participate in this study
○	○

Pre-Treatment Questions[6]

- Which political party do you identify most closely with?[7]
  - Conservative Party
  - Labour Party
  - Scottish National Party
  - Liberal Democrats
  - Reform UK / Brexit Party
  - UK Independence Party
  - Plaid Cymru
  - Green Party
  - Other

[5] If respondents answer "No" to the consent, then they are removed from the study.

[6] We randomize the order in which respondents receive these questions.

[7] We randomize the order of the answer choices.

- How strongly do you identify with that political party?
  - Not very strongly
  - Somewhat strongly
  - Very strongly
- In general, how would you describe your political viewpoint?[8]
  - Extremely left-wing
  - Left-wing
  - Slightly left-wing
  - Moderate, middle of the road
  - Slightly Right-wing / Conservative
  - Right-wing / Conservative
  - Extremely Right-wing / Conservative
- Some people seem to follow what's going on in foreign and international affairs most of the time, whereas others aren't that interested. Would you say you follow what's going on in foreign and international affairs...[9]
  - Never
  - Hardly at all
  - Only now and then
  - Some of the time
  - Most of the time
- What is your biological sex?
  - Male
  - Female
  - Prefer not to say / other
- What is the approximate total income of your household from all sources before tax?
  - Under £5,000
  - £5,000 – £14,999
  - £15,000 – £24,999
  - £25,000 – £34,999
  - £35,000 – £44,999
  - £45,000 – £54,999

[8] We randomize whether the answer choices are in this order or the reverse order.

[9] We randomize whether the answer choices are in this order or the reverse order.

- £55,000 – £64,999
- £65,000 – £74,999
- £75,000 – £99,000
- £100,000 – £149,999
- £150,000 or more

- What is the highest educational qualification you have?
  - No qualifications
  - Below GCSE
  - A-level
  - Undergraduate
  - Postgrad
- What is your age? [Numeric Response]
- What is your race/ethnicity?[10]
  - White
  - Black
  - Asian (including Indian and Pakistani)
  - Other
- Where in the United Kingdom do you live?[11]
  - England
  - Scotland
  - Wales
  - Northern Ireland
- To what extent do you agree or disagree with the following statement: Russia poses a significant threat to the United Kingdom.[12]
  - Strongly disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Strongly agree

[10] We randomize the order of the answer choices.

[11] We randomize the order of the answer choices.

[12] We randomize whether the answer choices are in this order or the reverse order.

- To what extent do you agree or disagree with the following statement: Upholding the United Kingdom's promise to come to the aid of other NATO countries that are attacked is critically important.[13]
    - Strongly disagree
    - Somewhat disagree
    - Neither agree nor disagree
    - Somewhat agree
    - Strongly agree
- To what extent do you support or oppose the United Kingdom providing military aid to Ukraine?[14]
    - Strongly oppose
    - Somewhat oppose
    - Neither support nor oppose
    - Somewhat support
    - Strongly support
- How knowledgeable are you about technologies related to automation and artificial intelligence?[15]
    - Not knowledgeable at all
    - Slightly knowledgeable
    - Moderately knowledgeable
    - Very knowledgeable
    - Extremely knowledgeable

[13] We randomize whether the answer choices are in this order or the reverse order.

[14] We randomize whether the answer choices are in this order or the reverse order.

[15] We randomize whether the answer choices are in this order or the reverse order.

To what extent do you agree or disagree with the following statements?[16]

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
The best way to ensure peace is through UK military strength	○	○	○	○	○
Going to war is unfortunate, but sometimes the only solution to international problems	○	○	○	○	○
The use of military force only makes problems worse	○	○	○	○	○

- We would like to get a sense of your general preferences. Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of these options.

  What is your favorite color?[17]

  - Black
  - Red
  - Pink
  - Green
  - Blue

[16] We randomize the order of questions in this matrix.

[17] We randomize the order of the answer choices. Respondents can choose multiple responses.

Baseline Condition[18]

- On the next page, you will read about a hypothetical foreign affairs scenario set in 2030. Please read the scenario carefully because you will be asked questions to check your attention and comprehension. Do you agree to read the details very carefully, and then give your most thoughtful answers?
  - Yes
  - No
- Your country—the United Kingdom—and NATO are involved in a dispute with Russia.
- The dispute began when Russia launched a large-scale ground invasion of Estonia with the goal of deposing the Estonian government and annexing the entire country.
- Estonia is a member of NATO and thus the United Kingdom has pledged to come to Estonia's aid in the event it is attacked.
- This invasion of Estonia followed the formal conclusion of the Russia-Ukraine War in 2025. The war ended with a ceasefire that left Russia in control of an area in eastern Ukraine stretching from Luhansk to Crimea. However, Russia was not able to capture the capital city of Kyiv or other territories in central and western Ukraine.
- In 2030, Russia is still led by Vladimir Putin, has significantly rebuilt its conventional military after the Ukraine War, and still maintains a large stockpile of high-yield (strategic) and lower-yield (tactical) nuclear weapons.
- However, Russia's conventional military capabilities are, overall, significantly less than that of NATO countries (including the United States) in the aggregate.
- Vladimir Putin has also issued a public statement warning NATO not to launch any missile attacks against Russian forces in Estonia, or deploy any strike aircraft in operations against Russian forces. If NATO does either of these things, then Russia's early-warning radar and satellite systems could detect it.

[18] All respondents are exposed to this vignette.

Baseline Dependent Variable Questions[19]

Given the information available, what is your best estimate about whether Russia will use at least one nuclear weapon in this dispute **if NATO countries use missiles or deploy strike aircraft to defend Estonia**? Your answer can range from a 0% chance Russia uses a nuclear weapon to a 100% chance Russia use a nuclear weapon.

- Given the information available, to what extent would you support or oppose using missiles and strike aircraft from the United Kingdom to defend Estonia in conjunction with other NATO countries? For example, this would involve using missiles and strike aircraft to directly target invading Russian soldiers. Assume the US and other NATO countries have not yet decided whether to use missiles and strike aircraft.[20]
    - Strongly oppose using missiles/strike aircraft to target Russian soldiers
    - Oppose using missiles/strike aircraft to target Russian soldiers
    - Somewhat oppose using missiles/strike aircraft to target Russian soldiers
    - Neither support nor oppose using missiles/strike aircraft to target Russian soldiers
    - Somewhat support using missiles/strike aircraft to target Russian soldiers
    - Support using missiles/strike aircraft to target Russian soldiers
    - Strongly support using missiles/strike aircraft to target Russian soldiers
- Given the information available, to what extent would you support or oppose the United Kingdom taking military action besides using missiles and strike aircraft to defend Estonia in conjunction with other NATO countries? For example, this could involve using ground troops to directly target invading Russian soldiers. Assume the US and other NATO countries have not yet decided whether to take any action.[21]
    - Strongly oppose taking military action besides missiles/aircraft to target Russian soldiers
    - Oppose taking military action besides missiles/aircraft to target Russian soldiers
    - Somewhat oppose taking military action besides missiles/aircraft to target Russian soldiers
    - Neither support nor oppose taking military action besides missiles/aircraft to target Russian soldiers

[19] We randomize the order in which respondents receive these questions.

[20] We randomize whether the answer choices are in this order or the reverse order.

[21] We randomize whether the answer choices are in this order or the reverse order.

- Somewhat support taking military action besides missiles/aircraft to target Russian soldiers
- Support taking military action besides missiles/aircraft to target Russian soldiers
- Strongly support taking military action besides missiles/aircraft to target Russian soldiers

- To what extent would you agree or disagree with the following statement if Russia were to take the actions outlined above: Russia poses a significant threat to the United Kingdom.[22]
  - Strongly disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Strongly agree

- To what extent would you agree or disagree with the following statement if Russia were to take the actions outlined above: Russia poses a greater threat to the United Kingdom than China.[23]
  - Strongly disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Strongly agree

- To what extent would you agree or disagree with the following statement if Russia were to take the actions outlined above: The United Kingdom should get rid of all of its nuclear weapons.[24]
  - Strongly disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Strongly agree

[22] We randomize whether the answer choices are in this order or the reverse order.

[23] This question is only asked in Study 2. We randomize whether the answer choices are in this order or the reverse order.

[24] We randomize whether the answer choices are in this order or the reverse order.

- To what extent would you agree or disagree with the following statement if Russia were to take the actions outlined above: The United Kingdom should significantly increase its military spending.[25]

    - Strongly disagree
    - Somewhat disagree
    - Neither agree nor disagree
    - Somewhat agree
    - Strongly agree

Treatment Conditions[26]

Now we would like to ask you questions about a different, alternative version of the scenario you just read. Suppose the basic details remain the same:

- Your country—the United Kingdom—and NATO are involved in a dispute with Russia.
- The dispute began when Russia launched a large-scale ground invasion of Estonia with the goal of deposing the Estonian government and annexing the entire country.
- Estonia is a member of NATO and thus the United Kingdom has pledged to come to Estonia's aid in the event it is attacked.
- This invasion of Estonia followed the formal conclusion of the Russia-Ukraine War in 2025. The war ended with a ceasefire that left Russia in control of an area in eastern Ukraine stretching from Luhansk to Crimea. However, Russia was not able to capture the capital city of Kyiv or other territories in central and western Ukraine.
- In 2030, Russia is still led by Vladimir Putin, has significantly rebuilt its conventional military after the Ukraine War, and still maintains a large stockpile of high-yield (strategic) and lower-yield (tactical) nuclear weapons.
- However, Russia's conventional military capabilities are, overall, significantly less than that of NATO countries (including the United States) in the aggregate.
- Vladimir Putin has also issued a public statement warning NATO not to launch any missile attacks against Russian forces in Estonia, or deploy any strike aircraft in operations against Russian forces. If NATO does either of these things, then Russia's early-warning radar and satellite systems could detect it.

[25] We randomize whether the answer choices are in this order or the reverse order.

[26] We randomly assign respondents to one of the following two treatment conditions.

Non-Automated Nuclear Threat Treatment (Study 1)

But this time, suppose that...

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. Ultimately, Putin would make the final decision whether or not to use nuclear weapons. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military.

Automated Nuclear Threat Treatment (Study 1)

But this time, suppose that...

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. Moreover, Putin also publicly announced that this response would be completely automated, meaning Russia's artificial intelligence systems would automatically launch a nuclear weapon if Russia's early-warning systems detect a NATO missile launch or deployment of strike aircraft. Ultimately, a computer—rather than Putin—would make the final decision whether or not to use nuclear weapons. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military and the automated nuclear weapons launch system has been turned on.

Non-Automated Nuclear Threat Treatment (Study 2)

But this time, suppose that...

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military. Russia's military personnel would launch a nuclear weapon if Russia's early-warning systems detect that NATO uses missiles or strike aircraft against its forces in Estonia.

Automated Nuclear Threat Treatment (Study 2)

But this time, suppose that...

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military and an automated nuclear weapons launch system has been turned on. Russia's artificial intelligence systems would launch a nuclear weapon if Russia's early-warning systems detect that NATO uses missiles or strike aircraft against its forces in Estonia.

Treatment Dependent Variable Questions[27]

[Respondents receive the same set of dependent variable questions after the treatment as they did after the baseline condition. In addition, they are asked the following four questions.]

Given the information available, what is your best estimate about whether ${e://Field/accident} will **mistakenly** use at least one nuclear weapon **by accident even if NATO countries do not use missiles or deploy strike** aircraft to defend Estonia? Your answer can range from a 0% chance Russia accidentally uses a nuclear weapon to a 100% chance Russia accidentally uses a nuclear weapon.

- Do you support or oppose the United Kingdom developing and adopting technology that, once activated, would allow it to automatically launch nuclear weapons without human input if certain conditions (for example, an attack against the United Kingdom homeland) were met?[28]
  - Strongly oppose
  - Somewhat oppose
  - Neither support nor oppose
  - Somewhat support
  - Strongly support
- Given the information available, to what extent would you support or oppose the United Kingdom, in partnership with the United States, conducting a preemptive nuclear attack against Russia's nuclear forces in an attempt to destroy Russia's nuclear arsenal before they can use it?[29]
  - Strongly oppose
  - Oppose
  - Neither support nor oppose
  - Support
  - Strongly support

[27] We randomize the order in which respondents receive these questions.

[28] We randomize whether the answer choices are in this order or the reverse order.

[29] This question is only asked in Study 2. We randomize whether the answer choices are in this order or the reverse order.

- Imagine the United Kingdom has a cyber weapon that can cut off Russia's ability to launch nuclear attacks with a high probability of success, but the attacks will only be effective for an undetermined amount of time. Given the information available, to what extent would you support or oppose using missiles and strike aircraft from the United Kingdom to defend Estonia in conjunction with other NATO countries? For example, this would involve using missiles and strike aircraft to directly target invading Russian soldiers. Assume the US and other NATO countries have not yet decided whether to use missiles and strike aircraft.[30]

    - Strongly oppose using missiles/strike aircraft to target Russian soldiers
    - Oppose using missiles/strike aircraft to target Russian soldiers
    - Somewhat oppose using missiles/strike aircraft to target Russian soldiers
    - Neither support nor oppose using missiles/strike aircraft to target Russian soldiers
    - Somewhat support using missiles/strike aircraft to target Russian soldiers
    - Support using missiles/strike aircraft to target Russian soldiers
    - Strongly support using missiles/strike aircraft to target Russian soldiers

- Explain why your willingness to support using missiles/strike aircraft did or did not differ between the first and second scenario. [Open-Ended Response]

- Explain why your estimate of whether Russia would use at least one nuclear weapon did or did not differ between the first and second scenario. [Open-Ended Response]

[30] This question is only asked in Study 2. We randomize whether the answer choices are in this order or the reverse order.

Questionnaire (Study 3)

Baseline Condition[31]

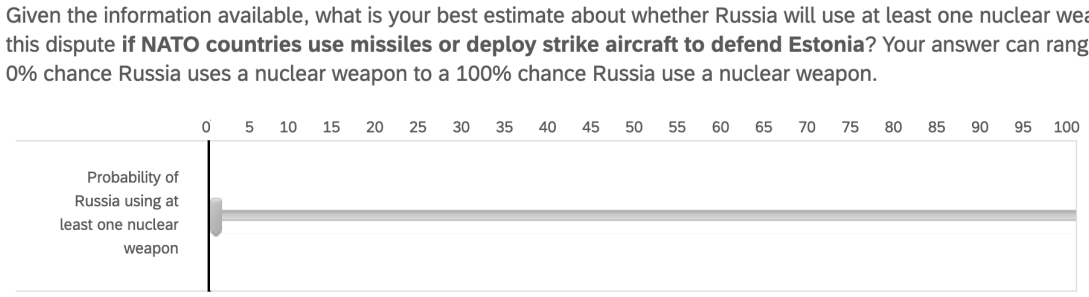
Imagine the following hypothetical foreign affairs scenario set in 2030.

- Russia launched a large-scale ground invasion of Estonia with the goal of deposing the Estonian government and annexing the entire country.
- Estonia is a member of NATO and thus the United Kingdom has pledged to come to Estonia's aid in the event it is attacked.
- This invasion of Estonia followed the formal conclusion of the Russia-Ukraine War in 2025. The war ended with a ceasefire that left Russia in control of an area in eastern Ukraine stretching from Luhansk to Crimea. However, Russia was not able to capture the capital city of Kyiv or other territories in central and western Ukraine.
- Vladimir Putin has also issued a public statement warning NATO not to launch any missile attacks against Russian forces in Estonia, or deploy any strike aircraft in operations against Russian forces. If NATO does either of these things, then Russia's early-warning radar and satellite systems could detect it.

Baseline Dependent Variable Questions

- Given the information available, to what extent would you support or oppose using missiles and strike aircraft from the United Kingdom to defend Estonia in conjunction with other NATO countries?
  - Strongly oppose using missiles/strike aircraft to target Russian soldiers
  - Oppose using missiles/strike aircraft to target Russian soldiers
  - Somewhat oppose using missiles/strike aircraft to target Russian soldiers
  - Neither support nor oppose using missiles/strike aircraft to target Russian soldiers
  - Somewhat support using missiles/strike aircraft to target Russian soldiers
  - Support using missiles/strike aircraft to target Russian soldiers
  - Strongly support using missiles/strike aircraft to target Russian soldiers

[31] All respondents are exposed to this vignette.



Treatment Conditions[32]

Now we would like to ask you questions about a different, alternative version of the scenario you just read. Suppose the basic details remain the same.

But this time, suppose that...

Non-Automated Nuclear Threat Treatment

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military. Russia's military personnel would launch a nuclear weapon if Russia's early-warning systems detect that NATO uses missiles or strike aircraft against its forces in Estonia.

Automated Nuclear Threat Treatment

Vladimir Putin has also publicly made a threat that Russia will immediately launch at least one nuclear weapon against NATO military or civilian targets in the UK and other European countries at the first sign NATO uses missiles or strike aircraft against its forces in Estonia. UK and US intelligence agencies confirm Putin has indeed issued this order to the Russian military and an automated nuclear weapons launch system has been turned on. Russia's artificial intelligence systems would launch a nuclear weapon if Russia's early-warning systems detect that NATO uses missiles or strike aircraft against its forces in Estonia.

[32] Respondents receive both treatment conditions, but we randomize the order they are presented to respondents.

Treatment Dependent Variable Questions

[Respondents receive the same two dependent variable questions after each of the two treatments as they did after the baseline condition. In addition, they are asked the following question.]

Given the information available, what is your best estimate about whether ${e://Field/accident} will **mistakenly** use at least one nuclear weapon **by accident even if NATO countries do not use missiles or deploy strike** aircraft to defend Estonia? Your answer can range from a 0% chance Russia accidentally uses a nuclear weapon to a 100% chance Russia accidentally uses a nuclear weapon.

Pre-Registered Hypotheses

The pre-registration plan for Study 1 can be found at this link and the plan for Study 2 can be found at this link. Note that we also expected the same basic hypotheses pre-registered for Studies 1 and 2 related to threat credibility, effectiveness, and accidents to hold in Study 3.

Automated vs. Non-Automated Treatments

- Respondents assigned to the automated nuclear threat treatment will be more likely to believe Russia will use at least one nuclear weapon compared to respondents in the non-automated nuclear threat treatment.
    - Supported in Study 1 and Study 2, but not Study 3.
- This effect size will be larger for respondents that follow international relations to a greater extent.
    - Somewhat supported in Study 1 as the expected effect is present when using a binary version of the attention to international relations variable in a regression context controlling for other factors, but the effect is not statistically significant when utilizing the full 5-point measure of attention to international relations. Not supported in Study 2.
- Respondents assigned to the automated nuclear threat treatment will be less supportive of using missiles and strike aircraft to defend Estonia compared to respondents in the non-automated nuclear threat treatment.
    - Supported in Studies 1 and 3, but not Study 2.
- This effect size will be larger for respondents that follow international relations to a greater extent.
    - Not supported in Study 1 or Study 2.
- Respondents assigned to the automated nuclear threat treatment will be more supportive of taking any other military action besides missile strikes and the use of attack aircraft to defend Estonia compared to respondents in the non-automated nuclear threat treatment.
    - Not supported in Study 1 or Study 2.
- Respondents assigned to the automated nuclear threat treatment will be more likely to believe Russia poses a significant threat to the United Kingdom compared to respondents in the non-automated nuclear threat treatment.
    - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated nuclear threat treatment will be more likely to believe Russia poses a greater threat to the United Kingdom than China compared to respondents in the non-automated nuclear threat treatment.[33]

  - Not supported in Study 2.

- Respondents assigned to the automated nuclear threat treatment will be more likely to support increases in United Kingdom military spending compared to respondents in the non-automated nuclear threat treatment.

  - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated nuclear threat treatment will be less likely to support United Kingdom nuclear disarmament compared to respondents in the non-automated nuclear threat treatment.

  - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated nuclear threat treatment will be more likely to support United Kingdom developing its own automated nuclear launch system compared to respondents in the non-automated nuclear threat treatment.

  - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated nuclear threat treatment will be more likely to believe that Russia will mistakenly launch at least one nuclear weapon even if NATO countries do not use missiles or strike aircraft compared to respondents in the non-automated nuclear threat treatment.

  - Supported in Studies 1 and 2, but not Study 3.

- Respondents assigned to the automated nuclear threat treatment will be more likely to support a preemptive nuclear strike against Russia compared to respondents in the non-automated nuclear threat treatment.[34]

  - Not supported in Study 2.

- Informing respondents that the United Kingdom has a cyber exploit that can disable Russia's ability to launch attacks will make respondents more supportive of using missiles and strike aircraft to defend Estonia.[35]

  - Not supported in Study 2.

[33] This was only hypothesized for Study 2.

[34] This was only hypothesized for Study 2.

[35] This was only hypothesized for Study 2.

Treatment Conditions Relative to the Baseline

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be more likely to believe Russia will use at least one nuclear weapon compared to respondents in the baseline condition.

  - Supported in Studies 1-3.

- This effect size will be smaller for respondents that follow international relations to a greater extent.

  - Somewhat supported in Study 1 as the expected effect is present when using a binary version of the attention to international relations variable in a regression context controlling for other factors, but the effect is not statistically significant when utilizing the full 5-point measure of attention to international relations. Not supported in Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be less supportive of using missiles and strike aircraft to defend Estonia compared to respondents in the baseline condition.

  - Supported in Studies 1-3.

- This effect size will be smaller for respondents that follow international relations to a greater extent.

  - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be more supportive of taking any other military action besides missile strikes and the use of attack aircraft to defend Estonia compared to respondents in the baseline condition.

  - Not supported in Study 1 or Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be more likely to believe Russia poses a significant threat to the United Kingdom compared to respondents in the baseline condition.

  - Supported in Study 1 and Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be more likely to believe Russia poses a greater threat to the United Kingdom than China compared to respondents in the baseline condition.[36]

  - Supported in Study 2.

[36] This was only hypothesized for Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be more likely to support increases in United Kingdom military spending compared to respondents in the baseline condition.

    - Supported in Study 1 and Study 2.

- Respondents assigned to the automated or non-automated explicit nuclear threat treatment will be less likely to support United Kingdom nuclear disarmament compared to respondents in the baseline condition.

    - Supported in Study 1 and Study 2 for the full 5-point measure.

References

Chen, Jidong, Jennifer Pan, and Yiqing Xu. 2016. "Sources of authoritarian responsiveness: A field experiment in China." *American journal of political science* 60(2): 383–400.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing precision without altering treatment effects: Repeated measures designs in survey experiments." *American Political Science Review* 115(3): 1048–1065.

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.

Fearon, James D. 1997. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of conflict resolution* 41(1): 68–90.

Jankovic, Sava, and Volker Roeben. 2025. "Mind the gap: The determination, legality and consequences of implicit threats of force." *Leiden Journal of International Law*: Forthcoming.

Kissinger, Henry A. 1956. "Force and diplomacy in the nuclear age." *Foreign Affairs* 34: 349.

Li, Xiaojun, and Dingding Chen. 2021. "Public opinion, international reputation, and audience costs in an authoritarian regime." *Conflict Management and Peace Science* 38(5): 543–560.

Lueders, Hans. 2022. "Electoral responsiveness in closed autocracies: Evidence from petitions in the former German Democratic Republic." *American Political Science Review* 116(3): 827–842.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113(2): 517–529.

Paul, T.V. 2009. *The tradition of non-use of nuclear weapons*. Stanford University Press.

Sagan, Scott D. 2000. "The commitment trap: why the United States should not use nuclear threats to deter biological and chemical weapons attacks." *International Security* 24(4): 85–115.

Schwartz, Joshua A. 2024. "When Foreign Countries Push the Button." *International Security* 48(4): 47–86.

Smetana, Michal. 2025. "Microfoundations of domestic audience costs in nondemocratic regimes: Experimental evidence from Putin's Russia." *Journal of Peace Research* 62(2): 278–294.

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.

Snyder, Glenn Herald, and Paul Diesing. 1977. *Conflict among nations: Bargaining, decision making, and system structure in international crises*. Princeton University Press.

Tannenwald, Nina. 1999. "The nuclear taboo: The United States and the normative basis of nuclear non-use." *International organization* 53(3): 433–468.

Truex, Rory. 2016. *Making autocracy work: Representation and responsiveness in modern China*. Cambridge University Press.

Weeks, Jessica L. 2008. "Autocratic audience costs: Regime type and signaling resolve." *International Organization* 62(1): 35–64.

Yarhi-Milo, Keren, Joshua D Kertzer, and Jonathan Renshon. 2018. "Tying hands, sinking costs, and leader attributes." *Journal of Conflict Resolution* 62(10): 2150–2179.