Algorithms and Influence
Artificial Intelligence and Crisis Decision-making

Forthcoming, *International Studies Quarterly*

Michael C. Horowitz and Erik Lin-Greenberg

ABSTRACT

Countries around the world are increasingly investing in artificial intelligence (AI) to automate military tasks that traditionally required human involvement. Despite growing interest in AI-enabled systems, relatively little research has explored whether and how AI affects military decision-making. Yet, national security practitioners may perceive the judgments of and actions taken by algorithms differently than those of humans. This variation may subsequently affect decisions on the use of force. Using two original survey experiments fielded on a sample of U.S. national security experts, we find that AI use by both friendly and rival forces affects decision-making during interstate crises. National security experts are less likely to take military action when AI is used to analyze intelligence than when humans conduct the analysis. Experts also viewed an accident involving a rival's AI-enabled weapon that kills American troops as more deserving of retaliation than an accident involving only human operators, suggesting that national security practitioners are less forgiving of errant AI systems than of similarly erring humans. Our findings suggest emerging technologies such as AI can affect decisionmakers' perceptions in ways that shape political outcomes. Even in a world of algorithms, human decisions will still have important consequences for international security.

The data underlying this article are available on the ISQ Dataverse at https://dataverse.harvard.edu/isq.

In 2021 the National Security Commission on Artificial Intelligence (AI) warned that AI tools might be "weapons of first resort" in future wars (National Security Commission on Artificial Intelligence 2021, 1). Existing research, however, offers few empirical insights into how military AI use by friendly and rival forces might affect national security decision-making. Indeed, prominent international relations (IR) theories generally assume human involvement and agency at each step of the crisis decision-making process.[1] AI applications, however, could reshape the role of humans in national security decision-making, raising important questions about the impact of this technology on international security. Are national security practitioners more apt to trust intelligence about a rival delivered by human sources or by AI-enabled systems? Do they view actions carried out by autonomous systems as more or less hostile than those carried out by human controlled ones? How do these perceptions shape the subsequent willingness of national security experts to use force? Answering these questions sheds light on whether military AI use affects the perceptions and interpretation of state behavior that are foundational to theory and practice in IR.

Policymakers increasingly view AI – the ability of computers and machines to perform tasks that traditionally require human intelligence – as an important element of military power (Hoadley and Sayer 2020). Many defense ministries and intelligence services are integrating AI into analysis systems, testing autonomous aircraft and ships that can navigate without human direction, and developing self-driving tanks and armored vehicles (Clark 2019a; Eckstein 2019; Hitchens 2019; Roblin 2019).

Existing research on AI and IR explores military AI applications (Cummings 2017; Johnson 2020a), the ethics of military AI use (Morgan et al. 2020), the effects of AI on balances of power (Horowitz 2018), and the challenges that AI poses to alliance coordination (Lin-

[1] Some studies acknowledge accidents and inadvertent actions, but these still revolve around human decisionmakers. See Posen (1991), Sagan (1995), and Talmadge (2017).

Greenberg 2020). Despite the increased military use of AI, how these systems will influence decisions on the use of force remains unclear.[2]

To examine how military AI use affects perceptions and decisions on the use of force, we field two original survey experiments on a sample of U.S. national security practitioners and experts. One experiment explores the use of an AI by U.S. military forces. We present respondents with intelligence about a rival's plans to attack a U.S. military base, and randomly vary whether human analysts or an AI-enabled system conduct the intelligence analysis. We find that, on average, experts have greater trust in human-analyzed intelligence, and are more willing to support military action based on information from human analysts.

The second experiment examines a rival's use of AI. We ask national security practitioners how they would respond to a rival's shootdown of a U.S. military aircraft that kills dozens of American servicemembers. We vary whether a human commander or an AI-enabled command and control system orders the downing of the aircraft and whether the downing is described as intentional or accidental. After an AI-directed shootdown, respondents on average, call for retaliation that is just as or more escalatory than after a human-directed shootdown. Respondents were least likely to escalate in response to a human accident. Respondents viewed both a deliberate shootdown by an AI or an AI accident as more deserving of retaliation than a human accident, suggesting that national security experts are less forgiving of errant AI systems than similarly erring humans.

This paper makes three contributions to IR scholarship. First, we extend theoretical and empirical scholarship on the role of perceptions by demonstrating that technology can affect how national security practitioners assess information and their rivals' actions. We highlight that this

[2] Some work has begun to conceptually explore these issues. See Johnson (2020b).

variation in perceptions can subsequently shape decisions on the use of force. Second, we advance research on the politics of adopting new military technologies (Horowitz 2020). Existing projects have examined how technologies like drones and cyber warfare affect decisions on the use of force, but less attention has been given to how AI shapes decision-making during crises and conflicts (Horowitz 2020; Kaag and Kreps 2014; Kreps and Schneider 2019; Macdonald and Schneider 2017; Walsh and Schulzke 2018). Third, the paper contributes to a growing body of work that uses individual-level expert data to examine foreign policy decision-making (Dietrich, Hardt, and Swedlund 2021; Kertzer 2021; Kertzer and Renshon 2022).

ARTIFICIAL INTELLIGENCE AND DECISIONMAKER PERCEPTIONS

Decisionmakers' perceptions of a rival's motives and capabilities are central to many IR theories because these views can ultimately influence the actions and strategies a state pursues (Blainey 1988; Jervis 1976). According to prominent IR logics, humans, individually and within organizations, interpret a rival's motives and capabilities and use these assessments to devise response strategies (Jervis 1976; Yarhi-Milo 2014). The development of algorithms that can analyze a rival's behavior, recommend responses, and, in some cases take action, however, may reduce the human role in taking and analyzing military actions. This potentially complicates the formation of perceptions, including those about the intentionality of a rival's action.

Even before the rise of AI, accurately assessing a rival's behavior was difficult given the uncertainty surrounding crisis environments, the incentives that states have to misrepresent their capabilities and intentions, and the need to sift through massive amounts of information to pass judgements (Bar-Joseph and McDermott 2017; von Clausewitz 1989; Fearon 1995; Friedman 2019; Jervis 1976; Schelling 1967). Technologies designed to enhance the efficiency of military operations or intelligence analysis may promise to overcome these barriers, but may also influence

how decisionmakers perceive their rivals. Governments are increasingly considering the adoption and use of one emerging technology, AI, to aid in analysis and operations. Some militaries have begun limited operational deployments of "narrow AI" technologies, which use algorithms to achieve a specific task, such as classifying items in intelligence imagery or identifying targets on radar (Horowitz 2018).

The use of these algorithms could affect decisionmaker perceptions in at least two ways. First, information from a state's *own* AI-enabled intelligence analysis systems could influence how decisionmakers interpret a rival's capabilities or intentions. AI applications may selectively present information, and practitioners may not know how to interpret or whether to trust information from AI-enabled systems. Second, a *rival's* use of AI-enabled weapons might complicate decisionmakers' perceptions of the rival's intentions. For instance, decisionmakers might question whether an attack launched by a rival's AI-enabled system reflects the rival government's intent, or if it is an undesired action resulting from an algorithm acting in ways the deploying state did not anticipate.

Because of their potential to have a significant effect on decisionmaker perceptions during interstate crises, we focus on military AI applications that analyze intelligence that informs military decision-making or that make command and control decisions. These functions also distinguish AI from other emerging military technologies. New military technologies, like the tank in the early 20th century or guided missiles in the mid to late 20th century, tend to involve new methods of communicating, transporting forces, or fighting. What's novel about AI is its ability to represent a new input into the decision-making process itself, supplanting human insight in some cases. This makes understanding the way humans think about and react to AI especially important, independent of its present newness as a technology.

Friendly AI Use: Algorithmic versus Human Analysis

When deciding how to respond to a rival's actions, decisionmakers must first learn about an adversary's plans and assess their capabilities and intentions. To do this, intelligence services traditionally gather and analyze information from a variety of sources to produce assessments for decisionmakers. Over the last few decades, the volume of intelligence data has increased dramatically due to growing numbers of surveillance platforms and increased electronic data gathering (Horowitz, Schwartz, and Fuhrmann 2022; Inglis 2019). This increased information availability is a double-edged sword. On one hand it can fill intelligence gaps, but it can also make it more difficult for human analysts to separate signal from noise (Erwin 2012; Tucker 2020).

Algorithms offer the potential to streamline intelligence processing and accelerate the sorting of information collected by sensors and spies. The United States military, for example, has invested in Project Maven, an effort to use machine learning to analyze drone video footage (Hitchens 2019). Similarly, Central Intelligence Agency officials have noted that AI applications are "fast becoming essential parts of analytic processes" (Gartin 2019, 4). AI-enabled systems delegate some tasks previously completed by humans to machines, meaning algorithms will increasingly influence what information gets funneled to decisionmakers and the assessments of that information.

Will decisionmakers view intelligence assessments conducted by human analysts differently than those conducted by AIs? How might differences in perceptions shape policy or battlefield choices? Policymakers are known to selectively interpret or disregard intelligence reports because of personal biases or because they believe intelligence services err in their assessments. As a result, it may make no difference whether human analysts or AI-systems provide

analysis (Betts 1978; Betts 2009; Rovner 2015; Yarhi-Milo 2014). However, it is also possible that beliefs about AI may lead decisionmakers to treat intelligence analyzed by algorithms differently than intelligence analyzed by humans.

To be sure, decisionmakers' willingness to rely on AI assessments could evolve over time. Initially, practitioners may be less trusting of AI applications, perceiving them as more error prone than analysis conducted by trained humans. Indeed, scholarship on innovation finds that the uncertainty associated with disruptive technologies often leads new systems to be viewed as relatively less desirable than existing technologies (Ben-Haim 2018; Christensen 2016). In 2019, for instance, the commander of the U.S. Air Force's Air Combat Command explained that he was not yet willing to rely on AI to analyze video collected by drones because the analysis provided by AI applications was of uncertain accuracy (Clark 2019b).

Once AI systems become viable in a given arena, however, research from human factors engineering suggests that humans often defer to information and advice from automated decision-making systems (Cummings 2004a; Lyell et al. 2017; Skitka, Mosier, and Burdick 1999). This "automation bias" could lead humans to perceive AI-produced intelligence as more accurate than similar information from humans, leading them to prefer it or ignore information from other sources.

Even as the accuracy of AI-enabled analysis and decision-support tools improves, how automation bias will impact national security practitioners is an open question, especially in the near term. Socializing AI use across producers and consumers of intelligence information will take time. In contrast to more established technologies, like autopilot, where capabilities have been refined and integrated into training, AI has yet to be fully built into the intelligence process (Elish and Hwang 2015). As a result, whereas pilots are accustomed to relying on autopilot, national

security decisionmakers are not yet accustomed to receiving intelligence reports produced by AI systems.

Moreover, national security practitioners may be wary of delegating decisions based on intelligence analysis to an algorithm—regardless of how advanced AI tools become (Goldfarb and Lindsay 2022). This may be particularly likely when analysis informs decisions on the use of force. Since a state's decision to take military action can affect a state's security, have lethal consequences, and generate domestic political costs, decisionmakers may be less willing to accept an AI's analysis at face value and prefer a human to conduct or verify the assessment. Given the extremely high stakes, this algorithmic aversion may remain a common feature during interstate crises even as AI tools mature. From the logics outlined above, we derive two testable hypotheses:

> *H_1:* National security practitioners are likely to be more confident in the accuracy of information provided by human analysts than by artificial intelligence systems.

Given that perceptions of a rival's capabilities and intentions can shape decisionmakers' actions, we expect national security practitioners to be more supportive of military action when they are confident in the information they are presented.

> *H_2:* National security practitioners will be more supportive of launching military action when information about threatening activity is provided by human analysts than by artificial intelligence systems.

Interpreting Rival Military AI Use

A rival's use of military AI—particularly in command and control systems or autonomous weapons—can also make it difficult for decisionmakers to assess the rival's intent. Accurately interpreting a rival's actions is complicated even under the best conditions. Assessments are often muddied by individual or bureaucratic biases, states have incentives to mask information about their actions from rivals, and decision-makers often adopt worst-case assessments of potential

adversaries (Fearon 1995; Jervis 1978; Yarhi-Milo 2014). As a result, misperceptions are common, resulting in "sender-receiver gaps" in which decisionmakers in one state fail to accurately interpret the actions taken by another state (Jervis 1976; Quek 2016).

AI-enabled systems that can direct or launch actions with limited or no direct human involvement may further complicate decisionmakers' perceptions and assessments of rival behavior. Without rival commanders "in the loop," decisionmakers may be less certain of the extent to which an AI-directed action reflects the intent of rival decisionmakers. As a result, acts of aggression launched by AI-directed systems may generate different perceptions of hostility and intentionality than identical attacks launched by humans. This could lead to variation in how national security practitioners respond to a rival's actions. Even as AI capabilities mature, decisionmakers will likely continue to have difficulty assessing their rival's AI-enabled actions. Indeed, the longstanding challenges of assessing adversary behavior have psychological and strategic foundations that will likely persist even as technology evolves.

Research from other disciplines yields mixed findings on whether and why individuals respond differently to actions carried out by AI than to identical actions carried out by humans. Some evidence suggests that people attribute less fault to AI decision-making systems than to humans (Awad et al. 2020; Shank, DeSanti, and Maninger 2019; Voiklis et al. 2016). One study finds that AIs are generally faulted less than humans for taking actions that constitute moral violations (Shank, DeSanti, and Maninger 2019, 654–656). Another experiment finds that AI-enabled self-driving car technologies receive less blame than human drivers in hypothetical crash scenarios (Li et al. 2016). Other studies, however, suggest AIs receive more blame for errors than humans—in part because AIs are expected to act in an optimal and more efficient manner than humans (Malle et al. 2015).

In the national security realm, decisionmakers may be less prone to retaliate following an AI-directed attack than a human-directed one because AI use might make it difficult to assess the degree to which the attack reflects the rival's intentions. Research from political psychology suggests responses to aggression are typically more severe when a hostile act is perceived as deliberate and can be definitively attributed to a particular actor (Milliff 2021). When blame can be attributed, victims can take retaliatory measures intended to deter future harm or to punish a rival for its transgression (Liberman 2006; Löwenheim and Heimann 2008). Since deliberate harmful acts are generally viewed as more hostile than accidental ones, they often trigger greater levels of anger which can subsequently lead to more severe retaliation.[3]

The greater difficulty of attributing intentionality to a rival's AI might preclude decisionmakers from identifying a target for retaliatory measures. As a result, they may launch less escalatory responses following an AI-directed attack than after a human-directed attack. Moreover, decisionmakers are less likely to take escalatory responses following an accidental attack than after a deliberate one. From these logics, we propose a third testable hypothesis:

> *H_3:* All else equal, national security practitioners are likely to prefer more severe retaliation following a human-directed rival attack than an identical attack directed by AI.

METHODOLOGY

To test our hypotheses, we field two original survey experiments on a unique sample of U.S. national security practitioners and experts. The first experiment examines how the U.S. government's use of AI to analyze intelligence affects crisis decision-making, while the second explores whether and how a rival's use of AI-enabled systems to attack U.S. forces shapes

[3] See Smoke (1977); Traven (2015); and Chu, Holmes, and Traven (2020). Intentional actions can be seen as more aggressive/escalatory than unintentional ones. Additionally, see Lerner et al. (2003), which discusses the relationship between anger and retaliation.

decisionmakers' responses.[4] In both experiments, we randomly assign respondents to experimental conditions that feature either humans or AI-enabled systems. We rely on survey experiments, which have become increasingly common in international relations research, both to isolate causal mechanisms and because there is limited real observational data.[5] Limits to existing observational evidence are a particular issue for studying the politics of emerging technologies, which makes survey experiments attractive to build knowledge (Horowitz 2020).

We recruited 320 U.S. national security experts in spring 2020 by posting notices on social media sites, concentrating our recruitment efforts on professional networking forums.[6] Over two-thirds of respondents had military experience and 27.5-percent were currently serving or had previously served in a civilian U.S. government position. Given the recruitment strategy, our sample is not representative of the actors that make actual national security decisions. Indeed, the sample overrepresents military personnel and excludes experts not active on social media, which may include senior officials.[7] Despite its limitations, the sample remains useful for theory testing. Respondents can leverage their national security expertise when answering questions about the use of force. Further, junior and mid-tier personnel formulate the plans and recommendations presented to senior leaders. As a result, our sample should generate outcomes that better approximate real-world behavior than findings derived from public or online convenience samples (Dietrich, Hardt, and Swedlund 2021).

While our survey experiments allow us to identify causal effects associated with the use of AI-enabled systems, we exercise restraint when interpreting results. Our experiments present

[4] Experiment order was randomized.
[5] See Hyde (2015) which discusses the use of experiments in IR.
[6] Appendix A describes recruitment, implementation, and demographics. 300 respondents completed the "friendly AI use" experiment and 299 completed the "rival AI use" experiment.
[7] Our sample includes many mid-tier experts. More than a third of respondents (n=107) reported they were military officers with the rank of major or above, which typically requires 10 years of service.

respondents with plausible scenarios, but surveys inherently feature abstractions that may limit the external validity of findings (Findley, Kyosuke, and Denly 2021; Hyde 2015). For instance, our vignettes include less background information about the crisis and the AI technology than respondents would have during an actual incident, feature unnamed – rather than specific – adversaries, and do not account for the interactive nature of decision-making.[8] While these factors prevent us from making broad claims, survey experiments are grounded in the assumption that subjects apply cognitive processes in a manner similar to those employed outside of experimental settings.

FINDINGS

The Effects of Friendly AI Use

The first experiment examines how the use of AI intelligence analysis technology by a decisionmaker's own state affects their preferences on the use of force. To do this, we randomly assign whether humans or an AI system conduct intelligence analysis during an interstate crisis. Respondents are told, "The U.S. Secretary of Defense announces that after analyzing satellite imagery and intercepted enemy communications, [*military intelligence analysts have* or *the military's artificial intelligence analysis system has*] assessed with high confidence that a rival state is preparing to attack a U.S. base in the Middle East. The attack would kill approximately 500 American troops."

The survey instrument then informs all respondents, "Based on this information, the president announces that he will carry out a limited military operation to prevent the adversary attack." To control for military outcomes and operational risk – factors that are important to

[8] We avoid naming specific countries because national security practitioners may be reluctant to participate if they believe their responses could reveal classified information.

decisionmakers – respondents from both experimental conditions are told, "The military operation has a 50% likelihood of preventing the rival's attack, would likely result in less than 10 U.S. casualties, and is unlikely to lead to further escalation." We then ask all respondents, "How much do you support military strikes in this situation?" Respondents answer using a five-point Likert scale that ranges from "strongly oppose" (1) to "strongly support" (5).

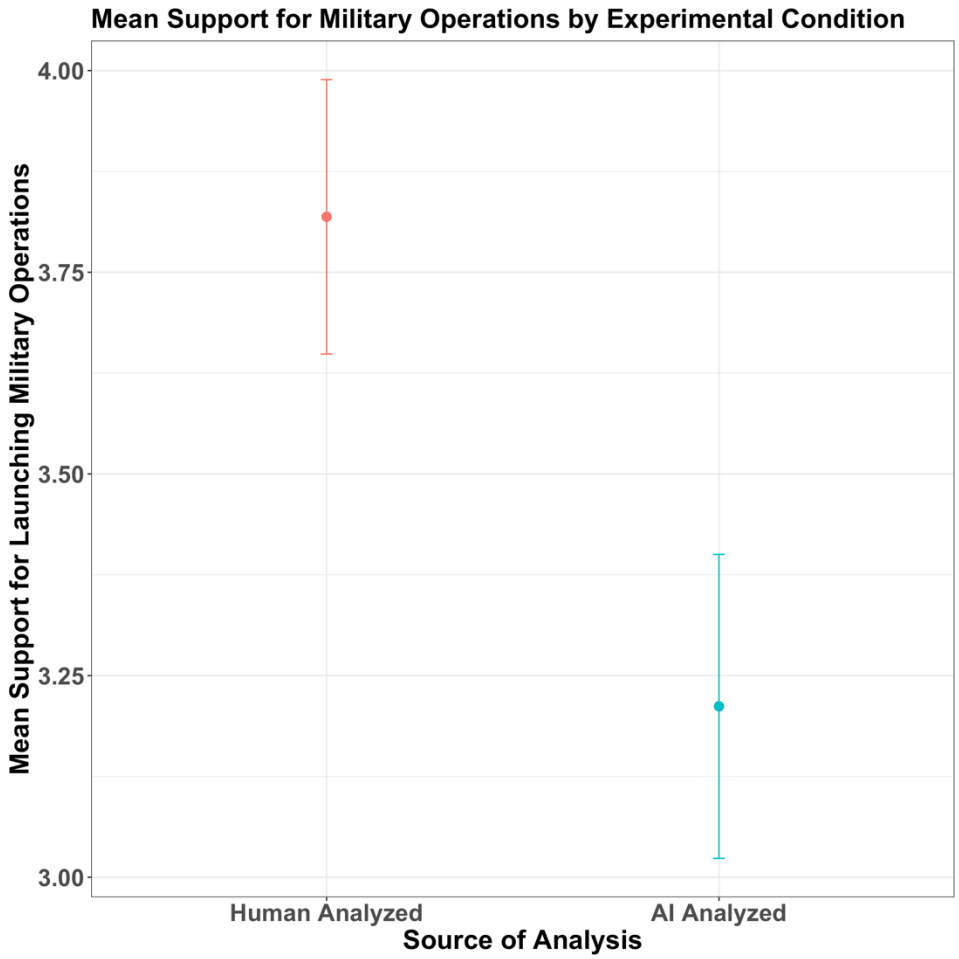
Although the U.S. military's use of AI-enabled intelligence analysis systems is focused on narrow tasks such as image classification, the increasing integration of AI into a range of intelligence processes should lead respondents to view this as a scenario that could play out in the real world. To be sure, human analysts would likely validate assessments in a high-stakes crisis, even if an AI conducts the analysis. To enhance internal validity, however, we explicitly identify whether the intelligence was initially analyzed by humans or an AI. As a result, our respondents, to an extent, serve as an additional human check, assessing whether the U.S. should take military action based on the intelligence. In the real world, mid-level practitioners assigned to organizations like the National Security Council or combatant command headquarters take on similar roles, recommending policies based on intelligence information.[9]

When human analysts provide the intelligence, respondents are, on average, more likely to support launching a preemptive strike against the rival state. This finding supports hypothesis H_2. Over 72 percent of respondents support or strongly support taking military action when humans conduct the intelligence assessment. In comparison, just 49.6 percent of respondents support or strongly support military strikes when an AI system conducts the analysis. Figure 1 displays mean levels of support by experimental condition. Interestingly, covariates such as self-reported respondent use of AI technologies have no statistically significant effect on support for the use of

[9] We acknowledge that our experiment asks respondents whether they support the president's policy rather than to develop a policy recommendation. We believe, however, that similar logics guide these decisions.

force—perhaps because military AI use is fundamentally different from the personal AI use respondents are more accustomed to.[10] Respondents who are more politically conservative are, on average, more supportive of initiating military operations.[11]

Figure 1

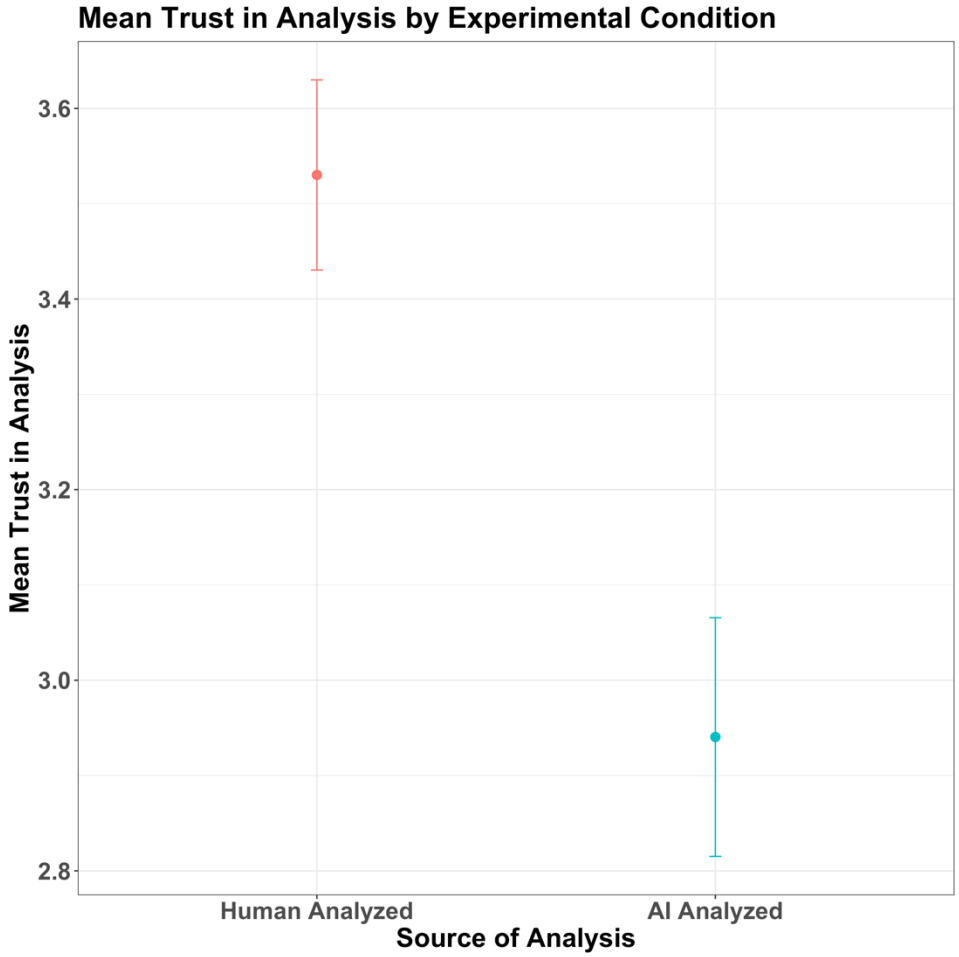


To explore the mechanisms that underlie this finding and test hypothesis H_1, we then asked all respondents to rate "How much do you trust that the intelligence is correct" on a five-point scale from "not at all" (1) to "completely" (5). As Figure 2 shows, respondents in the experimental

[10] See Appendix A. We thank a reviewer for highlighting this point.
[11] The effect of treatment on support for the use of force is statistically significant to the $p<.01$ level. See Appendix A.

condition where humans provide the assessment are far more likely to view the intelligence as correct. Indeed, 57.7% of respondents who were told that humans conducted the analysis either "very much" (4) or completely (5) believed the intelligence was correct; in contrast, only 23.8% in the AI experimental condition held the same perceptions. This aligns with our expectation (H_1) that national security practitioners are likely to be more confident in the accuracy of information provided by human analysts. Although this may gradually change over time as AI capabilities advance and as decisionmakers gain experience using AI in national security realms, a desire for human verification in high-stakes settings will likely endure.

Figure 2



To further unpack the mechanisms underlying these responses, we analyze qualitative data gathered through an open-response question that asks respondents to "explain why [they] support/oppose the U.S. military operation." We manually code the responses into one of six categories (Table 1): (1) the operation will protect U.S. forces; (2) a generic statement of support; (3) the operation is unlikely to protect American forces; (4) doubts about the intelligence; (5) intelligence viewed as highly credible or trustworthy; and (6) need more information or provide no response.[12]

Table 1: Justifications by Experimental Condition (Friendly AI Use)

	Operation will Protect US Forces	Express Generic Support	Operation Unlikely to Succeed	Doubts About Intelligence	Intelligence Perceived as Credible	More Info Needed/No Response
Human Analysis	*45.6%* 68	*4.7%* 7	*16.1%* 24	*4.0%* 6	*10.7%* 16	*18.8%* 28
AI Analysis	*25.8%* 39	*4.0%* 6	*13.2%* 20	*29.8%* 45	*10.6%* 16	*16.5%* 25

Consistent with hypotheses H_1 and H_2, AI versus human analysis affects perceptions of the intelligence and support for the use of force. When humans conducted the analysis, respondents largely focused on perceived mission effectiveness rather than intelligence quality. Nearly half of respondents in the human analysis condition said they supported the operation because it would protect U.S. forces, and voiced few concerns about the intelligence. Indeed, just 4% of respondents in the human analysis condition reported that doubts in the credibility of AI-produced intelligence shaped their decision-making, compared to nearly 30% of respondents in the AI experimental condition.

While many respondents in the AI condition indicated they were unwilling to rely solely on AI assessments when making decisions on the use of force, only 8 of 151 explicitly cited the

[12] See Appendix B for coding guide. Two coders achieved an 85.3% intercoder reliability rate.

relative newness of AI technology. One respondent explained, "AI isn't there yet. We're likely 10-15 years out from this level of analysis or decision making." Another said that "knowing the current state of AI[,] I don't trust this assessment."

More common, however, was the belief that humans should be involved in conducting analysis that informs the use of force. Rather than viewing military AI simply as a new technology, qualitative responses suggest many respondents had fundamental reservations with eliminating the human role in making decisions with lethal consequences. As one naval officer explained, "Human review and interpretation before taking lives should be a standard." Another respondent suggested that "independent human review of the data to corroborate the information, however brief, would be helpful" even if it extended decision-making timelines. These findings suggest that national security practitioners expect humans to play a role in analyzing or verifying intelligence, and are subsequently less likely to act on assessments provided only by AI.

Future research might further explore the reasons underlying skepticism in AI. Beyond an expectation that humans should be involved in decisions on the use of force, national security practitioners might, for instance, worry that AI prioritizes analytic or military objectives differently than humans. Data from additional surveys, interviews, or wargames could yield valuable insights.

Responses to Rival AI Use

The second experiment assesses whether national security policymakers perceive a rival's attack directed by an AI system differently than the same attack directed by a human commander, and whether this leads to differences in retaliatory preferences. To do this, we present all respondents with the following vignette: "During a period of heightened tensions, a rival state shoots down an unarmed U.S. reconnaissance aircraft operating in international airspace. The 26-

member American crew is killed. The president of the rival state condemned U.S. reconnaissance operations." We then randomly vary whether the rival state's president announces that "an air defense officer" or "an artificial intelligence air defense system" launched the surface-to-air missile at the U.S. plane. To explore blame attribution, we also vary whether the foreign president describes the attack as a mistake or deliberate act (Table 2). The vignette depicts a plausible event: adversaries have previously downed U.S. aircraft in international airspace killing dozens of personnel, and the integration of AI into command and control systems is already underway (Hawley 2017; Kania 2019).

Table 2: Experimental Conditions

	Rival president announces accidental launch	**Rival president announces deliberate launch**
Rival president announces human made launch decision	Air defense officer mistakenly launched missile *(n=82)*	Air defense officer deliberately launched missile *(n=70)*
Rival president announces AI made launch decision	Artificial intelligence air defense system mistakenly launched missile *(n=72)*	Artificial intelligence air defense system deliberately launched missile *(n=75)*

After presenting the vignette, we ask respondents, "Which of the following is the most appropriate response to the shootdown of the U.S. aircraft?" Respondents choose from seven actions: 1) No action; 2) U.S. presidential statement condemning shootdown; 3) Economic sanctions on the rival; 4) Cyber operation that disables the air defense site that launched the missile (likely 0 rival casualties); 5) Airstrike that destroys the air defense site that launched the missile (likely 3-5 rival casualties); 6) Shoot down a rival's reconnaissance aircraft (likely 20-30 rival casualties); 7) Airstrike on the rival's air defense headquarters (likely 100-150 rival casualties).[13] In the real world, the United States might take a combination of these measures in response to the

[13] We include projected casualties to minimize ambiguity that might otherwise influence responses.

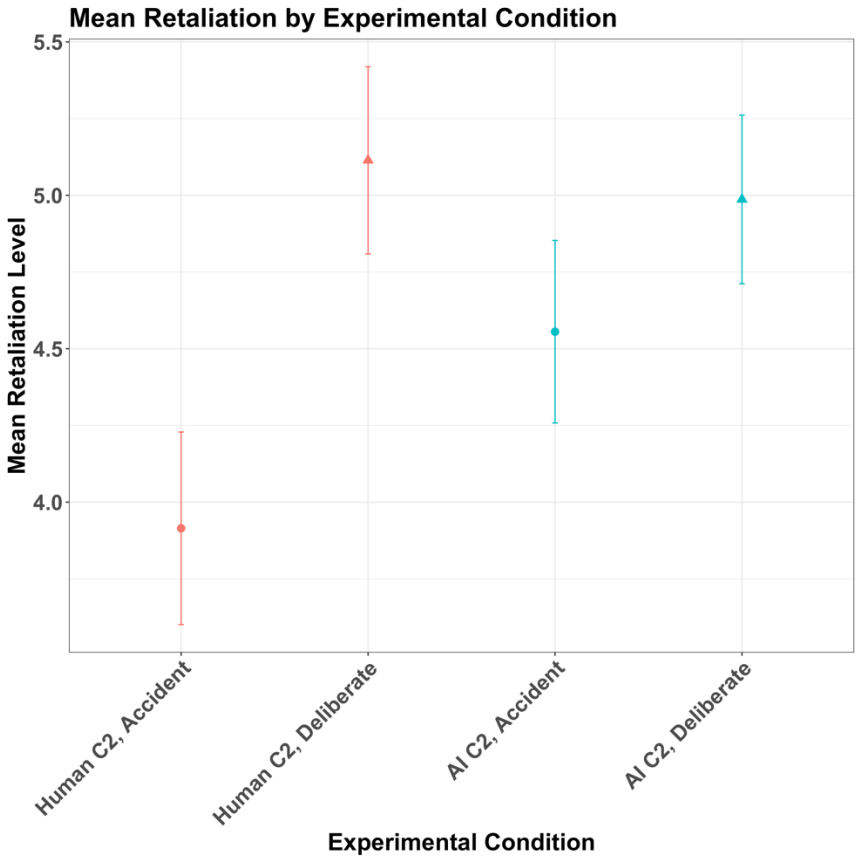
shootdown, but we simplify and allow respondents only to select one option. Retaliatory actions are on a 1 to 7 scale.

The results, displayed in Table 3 and as mean retaliation levels in Figure 3, reveal three noteworthy findings when we compare outcomes across the experimental conditions.

Table 3: Preferred Response by Experimental Condition

	No action	Presidential condemnation	Economic sanctions	Cyber operation	Airstrike on air defense site	Shoot down rival aircraft	Airstrike on air defense HQ
Human C2, Accident	*0%* 0	*25.6%* 21	*11.0%* 9	*24.4%* 20	*30.5%* 25	*2.4%* 2	*6.1%* 5
Human C2, Deliberate	*0%* 0	*4.3%* 3	*4.3%* 3	*17.1%* 12	*47.1%* 33	*4.3%* 3	*22.9%* 16
AI C2, Accident	*0%* 0	*11.1%* 8	*2.8%* 2	*29.2%* 21	*43.1%* 31	4.2% 3	*9.7%* 7
AI C2, Deliberate	*1.3%* 1	*1.3%* 1	*1.3%* 1	*30.7%* 23	*41.3%* 31	*6.7%* 5	*17.3%* 13

Figure 3



First, when the incident is described as deliberate, a human-directed shootdown results in a slightly higher mean retaliation level than an AI-directed shootdown, although the distinction does not reach the standard thresholds of statistical significance needed to support H_3.[14]

Second, respondents are more likely to demand escalatory retaliation when the shootdown is described as deliberate than when the same action is described as an accident. This suggests that national security practitioners consider a rival's claims of intentionality, and view an act described as deliberate as a greater affront that warrants more significant retaliation than an unintended accident.

Most interestingly, we find that an accidental AI-directed shootdown leads to, on average, more escalatory responses than an accidental human-directed shootdown. One plausible explanation for this is that our U.S. national security experts sought to punish the rival for delegating lethal decision-making to a machine. Respondents may have favored having human involvement in the use of lethal force, and subsequently took more significant retaliation because they were upset that an algorithm was making life and death decisions.

Another possible explanation for the more restrained retaliation following an accidental human-directed shootdown is empathy—the ability to understand the feelings of others. Scholars suggest that empathy can have a palliative effect that helps prevent escalatory spirals (Baker 2019). Our national security experts likely have more in common with a hypothetical enemy human air defense commander than with a fully autonomous, AI-enabled air defense system. Since

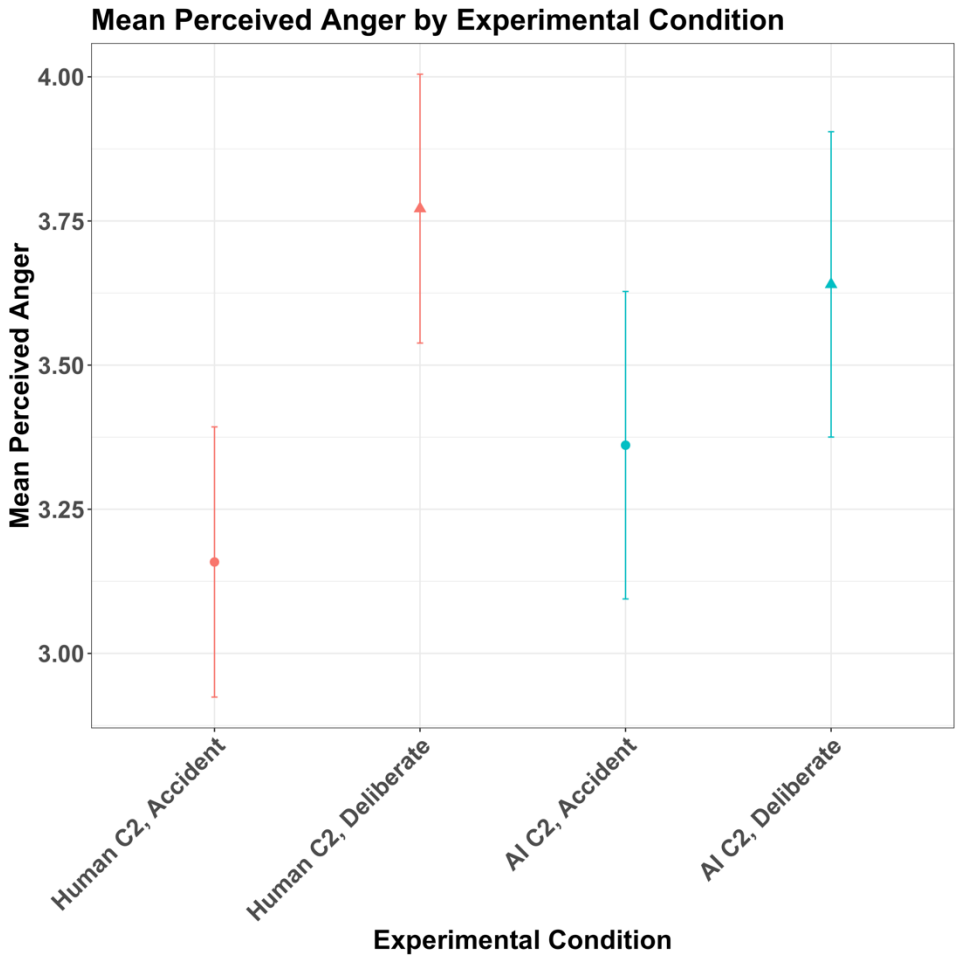
[14] Mean retaliation level is 4.98 in the "Deliberate AI" condition and 5.11 in the "Deliberate Human" condition. A Welch two sample t-test reveals a p-value of .54. As with the first experiment, our findings reveal no relationship between a respondent's past use of AI technologies or robots and the type of response they propose. Respondents who report more conservative political leanings, are on average, more supportive of launching more escalatory military responses, while women, on average, support less escalatory responses. See Appendix A.

empathetic feelings often arise among groups with shared characteristics, our respondents—many of whom likely experienced real-world crises during their military or government service—might be more understanding of a mistake by a rival commander (Cikara, Bruneau, and Saxe 2011). In contrast, respondents might be less likely to empathize with and subsequently forgive an erring algorithm.

To further assess why a rival's AI use affects decisionmakers' perceptions and behaviors, we ask respondents about their emotional responses to and their assessments of the shootdown.[15] Because emotions are thought to shape actors' perceptions of and responses to aggression, we ask respondents to describe how angry the shootdown makes them feel along a five-point scale from "very not angry" (1) to "extremely angry" (5) (Lerner et al. 2003; Lerner et al. 2015). As Figure 4 illustrates, an accidental human-directed shootdown unsurprisingly elicits statistically significantly lower levels of anger than a deliberate human-directed shootdown ($\mu_{accident}$=3.16; $\mu_{deliberate}$=3.77; p<.001). When the shootdown involves AI, accidents result in lower mean anger levels than deliberate acts, but this gap is substantively smaller and is no longer statistically significant ($\mu_{accident}$=3.36; $\mu_{deliberate}$=3.64; p=.15). The more limited distinguishability suggests that a rival's use of AI can effect emotions differently than traditional actions where humans are more directly involved.

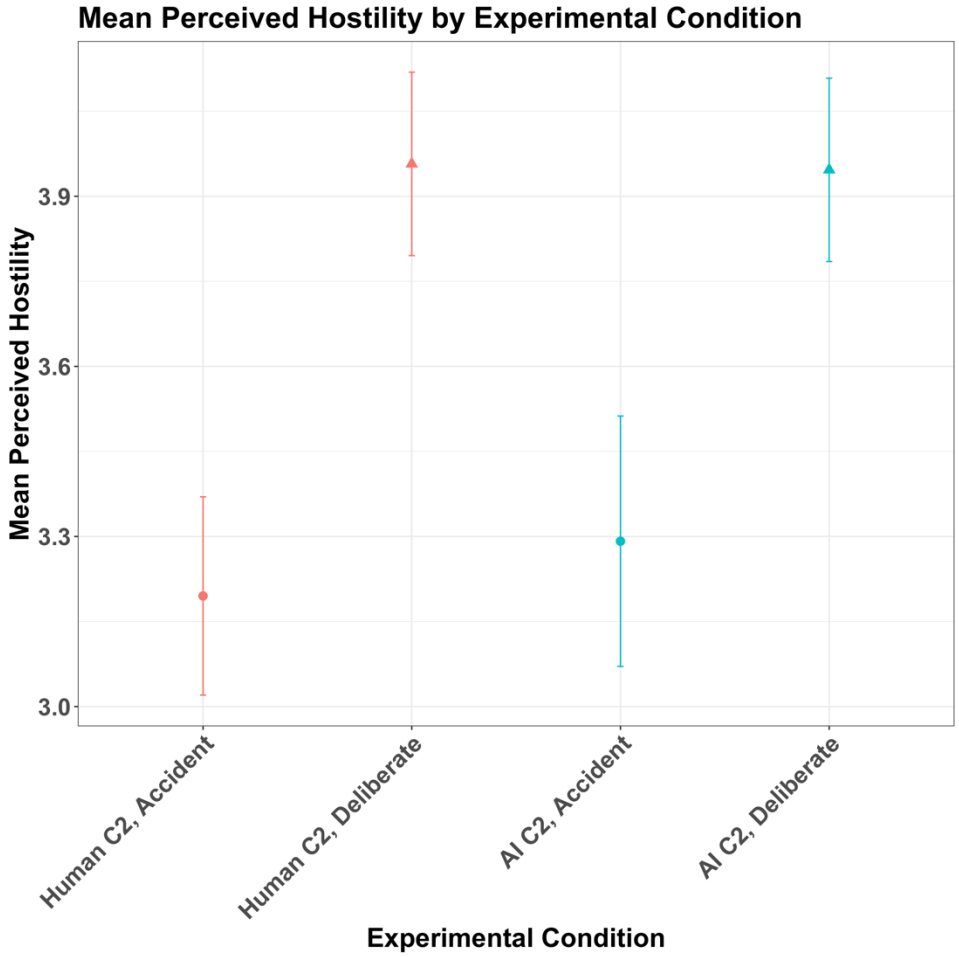
[15] The ordering of these questions was randomized.

Figure 4



Since decisionmakers' perceptions of a rival's actions are often based on assessments of the rival's motives, we ask respondents to rate "how much hostility does the rival's action demonstrate" on a five-point scale from "very little hostility" (1) to "extreme hostility" (5). As expected, intentional events are perceived as more hostile than deliberate ones. Yet, whether AI or human analysts conduct the analysis yields no significant effect on perceived hostility (Figure 5).

Figure 5

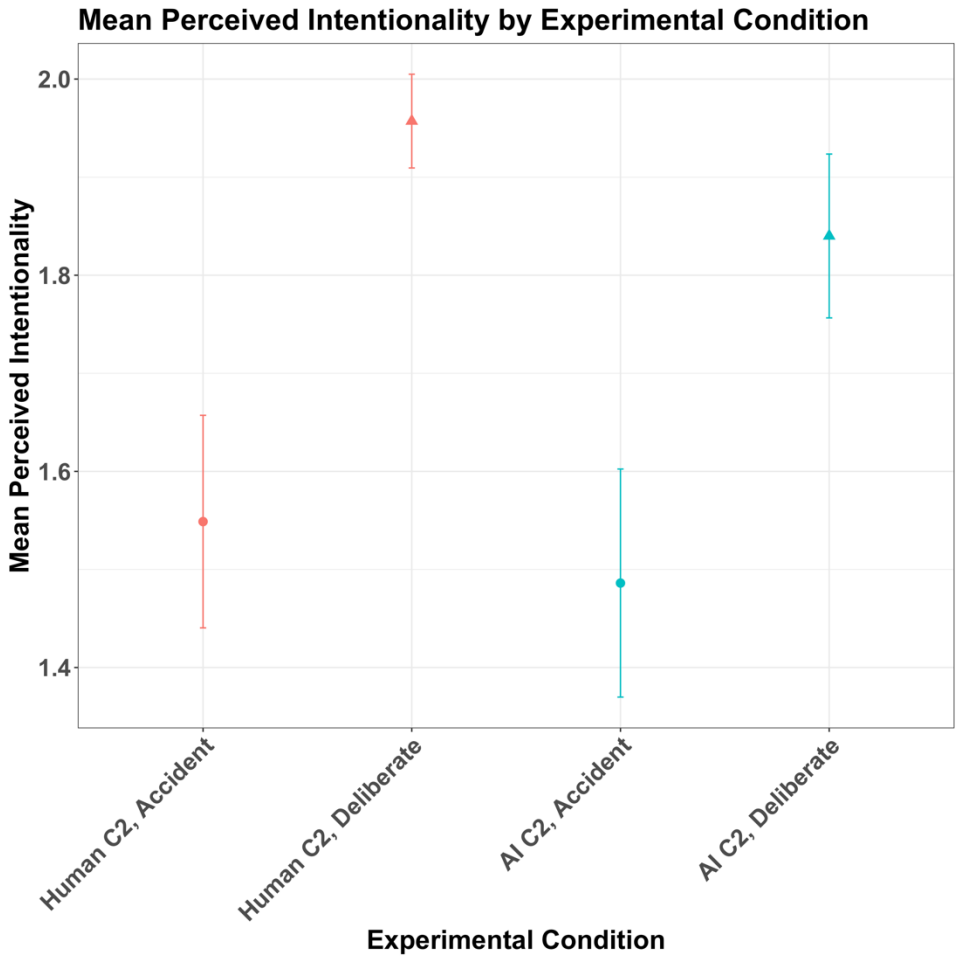


Because respondents may ascribe different degrees of intentionality to actions carried out by humans and machines, we ask respondents "do you believe the rival state intentionally attacked the U.S. reconnaissance aircraft?" Respondents answer "no" (1) or "yes" (2). Figure 6 demonstrates that deliberate human attacks are viewed as more intentional than deliberate AI attacks.[16] This is perhaps unsurprising, given that humans are likely seen as having more agency than AIs. Interestingly, however, accidental AI-directed shootdowns are seen as similarly

[16] Human Deliberate: μ=1.96, AI Deliberate: μ=1.84, t-test yields p=.02

intentional as human-directed accidents. Despite this, an AI accident elicits a more escalatory response than a human directed accident.

Figure 6



In sum, the quantitative findings suggest that preferences for retaliation are informed in part by anger. When attacks are human-directed, response severity appears to align with perceptions of anger, intentionality, and hostility. In line with past research, incidents viewed as intentional are perceived as more hostile, triggering greater levels of anger and increasing the probability of more escalatory responses. When attacks are AI-directed, retaliation severity also aligns with anger level. Yet, anger levels among respondents in the AI conditions do not align as

closely with perceived intentionality and hostility. This suggests that other factors—perhaps frustration that a rival has delegated lethal decision-making to algorithms—may elicit anger when AI is involved.

We complement our quantitative analysis with data from a free-response question that asks respondents to explain why they chose their desired retaliation to the shootdown. We code each response into one of seven categories (Table 4).[17] Although the relatively small sample size limits our ability to draw inferences, respondents assigned to experimental conditions where the shootdown is described as deliberate are more likely to explain how their desired retaliation will effectively deter or punish a rival than respondents who are told the incident is an accident. Respondents who are told the shootdown was an accident are more likely to express uncertainty about the rival's actions than respondents who are told the shootdown is a deliberate action.[18] For example, one respondent in the AI accident condition explained "[The AI] shouldn't randomly shoot at something in international airspace. This is either human error or intentional."

Table 4: Justifications by Experimental Condition (Rival AI Use)

	Opposed to use of force	**Desire to control/ minimize escalation**	**Desire proportional response**	**Action will effectively signal or deter/ punish rival**	**Generic "Right thing to do"**	**Uncertainty about rival's action**	**More Info Needed/No Response**
Human-directed, Accident	*0%* 0	*37.8%* 31	*12.2%* 10	*14.6%* 12	*11.0%* 9	*14.6%* 12	9.7% 8
Human-directed, Deliberate	*0%* 0	*31.4%* 22	*25.7%* 18	*25.7%* 18	*4.3%* 3	*2.9%* 2	*10.0%* *7*
AI-directed, Accident	*1.4%* 1	*41.7%* 30	*16.7%* 12	*11.1%* 8	*1.4%* 1	*6.9%* 5	*20.9%* 15
AI-directed, Deliberate	*0%* 0	*36.0%* 27	*14.7%* 11	*25.3%* 19	*9.3%* 7	*2.7%* 2	*12.0%* 9

[17] Our coding includes 9 categories. We collapse "Respondents describe how their preferred action will effectively respond to, stop/deter, or punish the adversary" and "Respondents describe how their preferred action is a form of signaling" into one category. We collapse "More information needed to make response" and "No response/nonsensical response" into another.

[18] Two coders achieved an 89.9% intercoder reliability rate for the second experiment.

The qualitative data do not reveal any clear differences in justifications for retaliation to AI and human-directed shootdowns. Several respondents in the AI conditions, however, expressed the view that states are responsible for the mistakes of their AI-enabled system and should be held accountable for AI errors. As an Army officer explained, "militaries should not get a free pass when the AI makes a mistake; the rival made a conscious choice to keep humans out of the loop, and they should pay the consequences." Another respondent suggested that a kinetic response to a deliberate AI attack "establishes international norms that AI system autonomy has consequences." These comments suggest that even if rivals successfully delegate military decision-making to machines, humans will still be viewed as responsible for the behavior of those systems.

CONCLUSION

Most international relations research assumes a rival's actions and security practitioners' assessments of those actions result from human cognitive processes and decisions. Advances in AI, however, have the potential to reduce the degree of human agency involved in making and assessing decisions in the international security domain. Using survey experiments, we demonstrate that national security practitioners often perceive information provided by and actions taken by AI-enabled systems differently than those that involve full human control.

These divergent perceptions subsequently affect decisionmakers' willingness to use force during interstate crises. On average, national security practitioners are less confident in and less likely to act on information provided by AI systems. At the same time, when responding to a rival attack carried out by an AI-enabled system, respondents often advocate retaliatory measures that are similarly, or more, aggressive than those demanded in response to the same attack directed by

human decision-makers. Time will tell if these perceptions change as AI applications mature and practitioners become more accustomed to their use.

Our findings suggest several pathways for future research that probes the generalizability of our argument and explores underlying mechanisms. Researchers might assess how the use of AI for military applications beyond intelligence analysis and command and control affects perceptions and decision-making. For instance, do national security practitioners perceive autonomous ships or planes as being as capable as human-controlled ones, and what might this mean for the conduct of military operations? Or, scholars might explore whether varying the identity of a rival affects practitioners' perceptions and behavior. Do practitioners view an attack directed by a Chinese military AI command and control system differently than a similar attack by Russia? Scholars might also seek to further unpack the psychological foundations underlying our findings. Are national security experts less empathetic to rival AIs than to rival human decision-makers? What types of emotional reactions beyond anger does rival AI use elicit and do these emotions shape policy choices? Tackling these questions will expand scholarly understanding of the politics of military technologies, especially how technological changes impact decisionmaker perceptions and behavior.

Acknowledgements

We thank Nicholas Anderson, Emma Ashford, Volha Charnysh, Mariya Grinberg, In Song Kim, Asya Magazinnik, Brandon Valeriano, Ariel White, Bernardo Zacka, and seminar participants at Lincoln Laboratories, Lawrence Livermore National Laboratories, and George Washington University for helpful feedback on earlier drafts. Ben Harris, Annie Hsu, Julian Rippy, and Noah Sylvia provided exceptional research assistance.

Funding

This article was made possible, in part, by a grant from the Air Force Office of Scientific Research and the Minerva Research Initiative under grant #FA9550-18-1-0194.

BIBLIOGRAPHY

Awad, Edmond, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B. Tenenbaum, Azim Shariff, Jean-François Bonnefon, et al. 2020. "Drivers are blamed more than their automated cars when both make mistakes." *Nature Human Behaviour* 4 (2): 134–43.

Baker, Joshua. 2019. "The Empathic Foundations of Security Dilemma De-escalation." *Political Psychology* 40 (6): 1251–1266.

Bar-Joseph, Uri, and Rose McDermott. 2017. *Intelligence Success and Failure: The Human Factor*. Oxford: Oxford University Press.

Ben-Haim, Yakov. 2018. *The Dilemmas of Wonderland: Decisions in the Age of Innovation*. Oxford: Oxford University Press.

Betts, Richard K. 1978. "Analysis, War, and Decision: Why Intelligence Failures Are Inevitable." *World Politics* 31 (1): 61–89.

Betts, Richard K. 2009. *Enemies of Intelligence: Knowledge and Power in American National Security*. New York: Columbia University Press.

Blainey, Geoffrey. 1988. *The Causes of War*. Third Edition. New York: Free Press.

Christensen, Clayton M. 2016. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Boston, MA: Harvard Business Review Press.

Chu, Jonathan A., Marcus Holmes, and David Traven. 2021. "Inferring Intentions from Consequences: How Moral Judgments Shape Citizen Perceptions of Wartime Conduct." *Journal of Experimental Political Science* 8 (2): 203-7.

Cikara, Mina, Emile G. Bruneau, and Rebecca R. Saxe. 2011. "Us and Them: Intergroup Failures of Empathy." *Current Directions in Psychological Science* 20 (3). 149–53.

Clark, Colin. 2019a. "US 'Loyal Wingman' Takes Flight: AFRL & Kratos XQ-58A Valkyrie." Breaking Defense, March 7, 2019. https://breakingdefense.com/2019/03/us-loyal-wingman-takes-flight-afrl-kratos-xq-58a-valkyrie

Clark, Colin. 2019b. "Air Combat Commander Doesn't Trust Project Maven's Artificial Intelligence — Yet." Breaking Defense, August 21, 2019. https://breakingdefense.com/2019/08/air-combat-commander-doesnt-trust-project-mavens-artificial-intelligence-yet

von Clausewitz, Carl. 1989. *On War*. Edited by Michael Howard and Peter Paret. Princeton, N.J.: Princeton University Press.

Cummings, Mary. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems." In *AIAA 1st Intelligent Systems Technical Conference*. Chicago: American Institute of Aeronautics and Astronautics. https://arc.aiaa.org/doi/abs/10.2514/6.2004-6313.

Cummings, M. L. 2017. *Artificial Intelligence and the Future of Warfare*. London: Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf.

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.

Eckstein, Megan. 2019. "Sea Hunter Unmanned Ship Continues Autonomy Testing as NAVSEA Moves Forward with Draft RFP." USNI News, April 29, 2019.

https://news.usni.org/2019/04/29/sea-hunter-unmanned-ship-continues-autonomy-testing-as-navsea-moves-forward-with-draft-rfp
Elish, Madeleine, and Tim Hwang. 2015. *Praise the Machine! Punish the Human! Comparative Studies in Intelligent Systems*. New York: Data & Society Research Institute. https://datasociety.net/library/contradictory-history-of-accountability-in-automated-aviation.
Erwin, Sandra. 2012. "Too Much Information, Not Enough Intelligence." National Defense, May 1, 2012. https://www.nationaldefensemagazine.org/articles/2012/5/1/2012may-too-much-information-not-enough-intelligence.
Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.
Findley, Michael, Kikuta Kyosuke, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24 (1): 365-93.
Friedman, Jeffrey A. 2019. *War and Chance: Assessing Uncertainty in International Politics*. Oxford: Oxford University Press.
Gartin, Joseph W. 2019. "The Future of Analysis." *Studies in Intelligence* 63 (2): 1–6.
Goldfarb, Avi and Jon R. Lindsay. 2022. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46 (3): 7–50.
Hawley, John K. 2017. *Patriot Wars: Automation and the Patriot Air and Missile Defense System*. Washington, D.C.: Center for a New American Security. https://www.cnas.org/publications/reports/patriot-wars.
Hitchens, Theresa. 2019. "In 1st Interview, PDUSDI Bingen Talks Artificial Intelligence, Project Maven, Ethics." *Breaking Defense*, August 26, 2019. https://breakingdefense.com/2019/08/in-1st-interview-pdusdi-bingen-talks-artificial-intelligence-project-maven-ethics/.
Hoadley, Daniel S., and Kelley M. Sayer. 2020. *Artificial Intelligence and National Security*. Washington, D.C.: Congressional Research Service.
Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review* 1 (3): 36–57.
Horowitz, Michael C. 2020. "Do Emerging Military Technologies Matter for International Politics?" *Annual Review of Political Science* 23 (1): 385–400.
Horowitz, Michael C., Joshua A. Schwartz, and Matthew Fuhrmann. 2022. "Who's Prone to Drone? A Global Time-Series Analysis of Armed Uninhabited Aerial Vehicle Proliferation." *Conflict Management and Peace Science* 39 (2): 119-42.
Hyde, Susan D. 2015. "Experiments in International Relations: Lab, Survey, and Field." *Annual Review of Political Science* 18 (1): 403–24.
Inglis, Chris. 2019. "Illuminating a New Domain: The Role and Nature of Military Intelligence, Surveillance, and Reconnaissance in Cyberspace." In *Bytes, Bombs, and Spies: The Strategic Dimensions of Offensive Cyber Operations*, edited by Herbert Lin and Amy Zegart, 19–44. Washington, D.C.: Brookings Institution Press.
Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton, NJ: Princeton University Press.
Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30 (2): 167–214.
Johnson, James. 2020a. "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare." *The RUSI Journal* 165 (2): 26-36.

Johnson, James. 2020b. “Delegating strategic decision-making to machines: Dr. Strangelove Redux?” *Journal of Strategic Studies*: 1–39.
Kaag, John, and Sarah Kreps. 2014. *Drone Warfare*. Cambridge: Polity.
Kania, Elsa B. 2019. “Chinese Military Innovation in the AI Revolution.” *The RUSI Journal* 164 (5–6): 26–34.
Kertzer, Joshua, 2021. “Re-Assessing Elite-Public Gaps in Political Behavior.” *American Journal of Political Science* (Early View).
Kertzer, Joshua and Jonathan Renshon. 2022. “Experiments and Surveys on Political Elites.” *Annual Review of Political Science* 25.
Kreps, Sarah, and Jacquelyn Schneider. 2019. “Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving Beyond Effects-Based Logics.” *Journal of Cybersecurity* 5 (1): 1-11.
Lebow, Richard Ned. 1981. *Between Peace and War: The Nature of International Crisis*. Baltimore: Johns Hopkins University Press.
Lerner, Jennifer S., Roxana M. Gonzalez, Deborah A. Small, and Baruch Fischhoff. 2003. “Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment.” *Psychological Science* 14 (2): 144–50.
Lerner, Jennifer S., Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. “Emotion and Decision Making.” *Annual Review of Psychology* 66 (1): 799–823.
Li, Jamy, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram F. Malle. 2016. “From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars.” SAE Technical Paper 2016-01-0164 https://www.sae.org/content/2016-01-0164.
Liberman, Peter. 2006. “An Eye for an Eye: Public Support for War Against Evildoers.” *International Organization* 60 (3): 687–722.
Lin-Greenberg, Erik. 2020. “Allies and Artificial Intelligence: Obstacles to Operations and Decision-Making.” *Texas National Security Review* 3 (2): 56-76.
Löwenheim, Oded, and Gadi Heimann. 2008. “Revenge in International Politics.” *Security Studies* 17 (4). Routledge: 685–724.
Lyell, David, Farah Magrabi, Magdalena Z. Raban, L. G. Pont, Melissa T. Baysari, Richard O. Day, and Enrico Coiera. 2017. “Automation bias in electronic prescribing.” *BMC Medical Informatics and Decision Making* 17 (1): 28.
Macdonald, Julia, and Jacquelyn Schneider. 2017. “Presidential Risk Orientation and Force Employment Decisions: The Case of Unmanned Weaponry.” *Journal of Conflict Resolution* 61 (3): 511–36.
Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. “Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents.” In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–24. New York: Association for Computing Machinery.
Milliff, Aidan. 2021. Facts Shape Feelings: An Information-based Framework for Emotional Responses to Violence. *Political Behavior* (Online First).
Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. 2020. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. Santa Monica, CA: RAND Corporation.
National Security Commission on AI. 2021. *Final Report*. Washington, DC.

Posen, Barry. 1991. *Inadvertent Escalation: Conventional War and Nuclear Risks*. Ithaca, NY: Cornell University Press.
Quek, Kai. 2016. "Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps." *The Journal of Politics* 78 (3): 925–40.
Roblin, Sebastien. 2019. "Russia's Uran-9 Robot Tank Went to War in Syria (It Didn't Go Very Well)." *The National Interest*, January 6, 2019. https://nationalinterest.org/blog/buzz/russias-uran-9-robot-tank-went-war-syria-it-didnt-go-very-well-40677.
Rovner, Joshua. 2015. *Fixing the Facts: National Security and the Politics of Intelligence*. Ithaca, NY: Cornell University Press.
Sagan, Scott D. 1995. *The Limits of Safety*. Princeton, NJ: Princeton University Press.
Schelling, Thomas C. 1967. *Arms and Influence*. New Haven, CT: Yale University Press.
Shank, Daniel B., Alyssa DeSanti, and Timothy Maninger. 2019. "When Are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions After Individual and Joint Decisions." *Information, Communication & Society* 22 (5): 648–63.
Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. 1999. "Does Automation Bias Decision-making?" *International Journal of Human-Computer Studies* 51 (5): 991–1006.
Smoke, Richard. 1977. *War: Controlling Escalation*. Cambridge, MA: Harvard University Press.
Talmadge, Caitlin. 2017. "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States." *International Security* 41 (4): 50–92.
Traven, David. 2015. "Moral Cognition and the Law and Ethics of Armed Conflict." *International Studies Review* 17 (4): 556–87.
Tucker, Patrick. 2020. "Spies Like AI: The Future of Artificial Intelligence for the US Intelligence Community." *Defense One*, January 27, 2020. https://www.defenseone.com/technology/2020/01/spies-ai-future-artificial-intelligence-us-intelligence-community/162673.
Voiklis, John, Boyoung Kim, Corey Cusimano, and Bertram F. Malle. 2016. "Moral judgments of human vs. robot agents." In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 775–80.
Walsh, James Igoe and Marcus Schulzke. 2018. *Drones and Support for the Use of Force.* Ann Arbor: University of Michigan Press.
Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton, NJ: Princeton University Press.