Introducing ICBe: An Event Extraction Dataset from Narratives about International Crises

A Preprint

Rex W. Douglass *
University of California, San Diego

Thomas Leo Scherer
University of California, San Diego

J Andrés Gannon
Vanderbilt University

Erik Gartzke
University of California, San Diego

Jon Lindsay
Georgia Institute of Technology

Shannon Carcelli
University of Maryland

Jonathan Wilkenfeld
University of Maryland

David M. Quinn
University of Maryland

Catherine Aiken
Georgetown University

Jose Miguel Cabezas Navarro
Universidad Mayor

Neil Lund
University of Maryland

Egle Murauskaite
University of Maryland

Diana Partridge
University of Maryland

November 20, 2023

Abstract

How do international crises unfold? We conceptualize international relations as a strategic chess game between adversaries and develop a systematic way to measure pieces, moves, and gambits accurately and consistently over a hundred years of history. We introduce a new ontology and dataset of international events called ICBe based on a very high-quality corpus of narratives from the International Crisis Behavior (ICB) Project. We demonstrate that ICBe has higher coverage, recall, and precision than existing state of the art datasets and conduct two detailed case studies of the Cuban Missile Crisis (1962) and the Crimea-Donbas Crisis (2014). We further introduce two new event visualizations (event iconography and crisis maps), an automated benchmark for measuring event recall using natural language processing (synthetic narratives), and an ontology reconstruction task for objectively measuring event precision. We make the data, supplementary appendix, replication material, and visualizations of every historical episode available at a companion website crisisevents.org.

*Correspondence should be addressed to Rex W. Douglass at rexdouglass@gmail.com.

If we could record every international interaction in the realms of diplomacy, conflict, economics, and beyond, how much unique information would this chronicle amount to, and how surprised would we be to see something new? In other words, what is the entropy of international relations? While this record could in principle be unbounded, the central conceit of social science is that there are structural regularities that limit what actors can do, their best options, and even which actors are likely to survive (Brecher 1999; Reiter 2015). If so, then these events can be recorded and systematically measured by social scientists interested in these regularities.[2] Thanks to improvements in natural language processing, more open-ended efforts have begun to capture entire unstructured streams of events from international news reports.[3] This has invited fruitful efforts to evaluate the coverage, quality, and accuracy of attempts to measure international affairs.

We advance existing efforts to identify and structure regularized events and actor in international politics by combining human coding with natural language processing to develop (1) a large, flexible ontology of international affairs and (2) a fine-grained and structured event dataset of international crises from 1918-2017 developed by applying our ontology to an unusually high-quality corpus of historical narratives of international crises (Brecher 1999; Wilkenfeld and Brecher 2000; Brecher et al. 2021). We then develop several methods for objectively gauging how well these event codings reconstruct the information contained in the original crisis narrative. We conclude by benchmarking our event codings against several current state of the art event data collection efforts. The underlying fine-grained variation in international affairs is unrecognizable through the lens of current quantification efforts. We find that existing models produce data on historical episodes that do not contain enough information to reconstruct the underlying event. In focusing this initial effort on international crises as a proof of concept sample, we demonstrate our ontology and method's potential to improve upon existing empirical identifications of patterns of international interactions.

In the proceeding four sections, this measurement paper makes the following arguments. First, there is a real-world unobserved latent concept known as international relations that can and should be systematically measured. Second, we propose a method for systematic large-scale measurement of the actors and behaviors in international affairs and as a proof of concept apply that method to a well-regarded and salient sample of events known as international crises. Third, in doing so we confirm that those measurements exhibit several desirable kinds of internal and external validity and out-perform existing approaches. Fourth, this validation can be evaluated in detail via new event visualizations, with examples provided for case studies of the 1962 Cuban Missile Crisis and 2014 Crimea-Donbas crisis. A final section concludes.

[2]See work on crises (Brecher and Wilkenfeld 1982; Beardsley et al. 2020), militarized disputes (Palmer et al. 2021; Gibler 2018), wars (Sarkees and Wayman 2010; Reiter, Stam, and Horowitz 2016), organized violence (Ralph Sundberg and Mihai Croicu 2016; Davies, Pettersson, and Öberg 2022), political violence (Raleigh et al. 2010), sanctions (Felbermayr et al. 2020), and international agreements (Kinne 2020; Owsiak, Cuttner, and Buck 2018), dispute resolution (Frederick, Hensel, and Macaulay 2017), and diplomacy (Moyer, Turner, and Meisel 2020; Sechser 2011).

[3]See Beieler et al. (2016); Boschee et al. (2015); Brandt et al. (2018); Grant et al. (2017); Li et al. (2021). On event extraction from images and social media see Zhang and Pan (2019) and Steinert-Threlkeld (2019).

1 Identifying and measuring international relations

1.1 Motivation

How can scholars abstract and measure discrete events about a historical episode in international relations? We employ a metaphor of chess, a game that despite its complexity can nonetheless be recorded in a standardized and structured manner. Like chess, international interactions involve a finite set of players (e.g. polities, rebel groups, IGOs) expressing their behaviors (e.g. thinking, saying, and doing) by moving pieces (e.g. military platforms, civilian personnel, diplomats) from and onto identifiable locations (geo-coded coordinates). These moves occur over a marked time period (start and end of a historical episode) and gambits occur in recordable sequences (actions and reactions) to produce observable, if still disputable, outcomes (e.g. victory, defeat, stalemate, peace). Our knowledge of this historical episode, including the actors involved as well as their preferences, behaviors, and beliefs are only indirectly observed from historical records that most often take the form of unstructured natural language text.

Much like the recording of chess evolved from natural language descriptive text notation to the modern figurine algebra notion, international relations scholars have recently sought to produce a structured account of historical events by combining the unstructured corpus of historical records with informative priors about international relations. The resulting structured account of information about a historical episode can be combined with that of other events to produce a systematic account of political events from which we can garner novel insight, assuming the underlying data have high coverage, precision, and recall. The easiest way to convey the desired produced of this task is with an example. Figure 1 shows a narrative account of the Cuban Missile Crisis (1962) in natural language sentences alongside a mapping to discrete machine-readable abstractive events. From the structured data, scholars can identify similarities and differences across events concerning important concepts like when particular foreign policy actions deter versus inflame (Jervis 1978; Glaser 2000), when third parties mediate in interstate disputes (Haffar 2002; Quinn et al. 2006), and how actors try to communicate resolve (Trager 2016; Lupton 2018). Identifying patterns of international interactions is not just an inherently interesting enterprise; it is a necessary precondition to important efforts to predict where policymakers should turn their attention to improve global welfare (Ward et al. 2013; Beger, Morgan, and Ward 2021).

1.2 Existing state of the art measurements

We draw informative prior beliefs about the underlying process of international relations that we expect to govern behavior during historical episodes and their conversion to the historical record. We organize our prior beliefs along two overarching axes: existing efforts to identify the actors/actions of international relations and identifying a corpus that can be used to produce an ontology of the information we hope to recover. Table 1 describes these two axes as columns and rows, respectively.

S	Natural Language Sentences (ICB Corpus)	Machine-Readable Events (ICBe)
4	When the U.S. discovered the presence of Soviet military personnel in Cuba on 7 September 1962 it called up 150,000 reservists.	mobilization 100ks discover fact - deployment to area 100s;1ks
5	The Soviets mobilized on the 11th.	mobilization 1ks
8	The U.S. crisis was triggered on 16 October when the CIA presented to President Kennedy photographic evidence of the presence of Soviet missiles in Cuba.	discover fact;start of crisis - deployment to area fortify start of crisis - deployment to area
9	The U.S. responded with a decision on the 20th to blockade all offensive military equipment en route to Cuba.	blockade coastline 10ks blockade coastline
10	When this was announced on 22 October, a crisis was triggered for Cuba and the USSR. An urgent meeting of the UN Security Council was requested by both the U.S. and Cuba on the 22nd, and by the USSR the next day.	start of crisis - appeal - meeting
11	On the 23rd as well, the Soviets accused the United States of violating the UN Charter and announced an alert of its armed forces and those of the Warsaw Pact members.	raise in alert 100ks accuse - violate terms of treaty
12	That day Cuba responded by condemning the U.S. blockade and declaring its willingness to fight.	- coastline disapprove
13	A resolution was adopted on the 23rd by the OAS calling for the withdrawal of the missiles from Cuba and recommending that member-states take all measures, including the use of force, to ensure that the government of Cuba would not continue to receive military material.	end military cooperation demand - withdraw from area 1ks
14	On 24 October the Security Council adopted a resolution requesting the Secretary-General to confer with the parties.	express intent - mediation express intent - mediation
15	On that same day, U Thant began mediation by sending identical letters to Khrushchev and Kennedy which proposed that the Soviet Union and the United States enter into negotiations, during which period both the shipment of arms and the quarantine would be suspended.	mediation
16	Moscow's major response to the crisis was a letter from Khrushchev to Kennedy on 26 October offering the removal of Soviet offensive weapons from Cuba and the cessation of further shipments in exchange for an end to the U.S. quarantine and a U.S. assurance that it would not invade Cuba.	offer happens - withdraw from area -C- will happen end blockade offer happens - -C- will happen end blockade
17	The situation was exacerbated on the 27th when a U.S. U-2 surveillance plane was shot down.	border violation airspace 1s 1s
18	That day another Khrushchev letter was received in Washington offering the removal of Soviet missiles from Cuba in exchange for the removal of U.S. missiles from Turkey.	offer happens - withdraw from area -C- will happen withdraw from area offer happens - -C- will happen
19	U.S. mobilization and aerial reconnaissance flights were stepped up.	border violation;mobilization mobilization airspace
20	And on the 27th President Kennedy sent the Soviet premier an acceptance of the proposals contained in the letter of 26 October while making no reference to Khrushchev's second letter of the 27th.	accept -
21	The following day Khrushchev notified the U.S. government that he had ordered work on the missile sites in Cuba stopped.	accept -
23	At the same time he warned Washington that U-2 reconnaissance flights over Cuba must be stopped as well.	demand - withdraw from area airspace
24	The crisis continued at a lower level of intensity for several more weeks due to Cuban President Castro's demands concerning a U.S. pledge not to invade his country.	demand - de-mobilization;lower alert ;body of water;coastline 100ks
25	On 30 October U Thant began talks in Havana, and Kennedy agreed to lift the quarantine for the duration of the talks.	end blockade coastline 10ks accept - end blockade coastline
26	When Cuba rejected UN inspection, the U.S. resumed the quarantine and air surveillance.	reject - unspecified cooperation reject blockade airspace;coastline 10ks
27	The Kremlin sent Deputy Premier Anastas Mikoyan to Cuba on 2 November to try to persuade Castro to allow UN inspection.	discussion;meeting
28	When this proved unsuccessful, a U.S.-USSR agreement was reached on 7 November allowing U.S. inspection and interception of Soviet ships leaving Cuba and the photographing of the missiles.	sign formal agreement
29	The following day the superpowers negotiated the removal of the IL-28 bombers which Castro had claimed were Cuban property.	withdraw from area 100s sign formal agreement
30	Castro's agreement was conveyed to the U.S. on 20 November 1962, which terminated the Missile crisis for all three actors.	end of crisis
31	The U.S. naval quarantine was lifted immediately, but aerial surveillance continued until the agreement was completely carried out.	end blockade body of water;coastline 10ks

Figure 1: Cuban Missile Crisis (1962) - ICB Narrative vs. ICBe Events

The rows in Table 1 represent the types of information we expect to find in international relations and forms the basis for our proposed ontology. We created this ontology by performing of the corpus and identifying named entities and verbs mentioned in the text. To identify possible behaviors, we matched verbs to the most likely definition found in Wordnet (Miller 1995), tallied (SI Appendix 1.2), and then aggregated them into a smaller number of behaviors balancing conceptual detail with manageable sparsity for human coding, informed by existing conceptual ("Literature" column) and measurement research (discussed below). We used the ICB project actor level data to identify likely actors for each crisis and location options relative to each actor. For behavior, actor, and location, coders could write-in a value if the given options were insufficient. The codebook lists 11 behaviors added post-coding as coders flagged events that were not captured by the initial ontology (e.g. propaganda).

As we are not the first to attempt to measure international relations in a structured manner, the columns of Table 1 compare the ontological coverage of ICBe to existing state of the art systems in production and with global coverage. We choose these datasets and models as they represent frequently used and reputable efforts to structure and describe historical events of interest to scholars of international politics. The first column starts with our contribution, ICBe, alongside other event-level datasets including CAMEO dictionary lookup-based systems (Historical Phoenix (Althaus et al. 2019); ICEWS (Boschee et al. 2015); Terrier (Grant et al. 2017)), the Militarized Interstate Disputes Incidents dataset, the UCDP-GED dataset (Davies, Pettersson, and Öberg 2022; Sundberg and Melander 2013), and ACLED (Raleigh et al. 2010).[4] The final set of columns compares episode-level datasets beginning with the original ICB project (Brecher et al. 2021; Brecher and Wilkenfeld 1982; Beardsley et al. 2020); the Militarized Interstate Disputes dataset (Palmer et al. 2021; Gibler 2018), and the Correlates of War (Sarkees and Wayman 2010). We include episode-level datasets as they remain and common and trusted tool for analyzing international relations, and because ICBe is unique among event-level datasets as events are matched to crises and so can be aggregated to the episode level. There is imperfect overlap concerning their intended depth and scope of coverage; 'international crises' are similar, but not identical to, 'interstate wars' and 'militarized interstate disputes' which differ yet again from 'individual events of organized violence' and 'non-violent action'. Even like-concepts require care in comparison, as an 'aim' in ICBe is the same as in MIPS, but an 'alert' in ICBe is not the same as an 'alert' in MID.

This comparison is not intended to fault existing data and models for not including every variable in ICBe's ontology, as some of these variables fall outside the scope of a particular dataset's intended purpose. Rather, it serves as an initial basis for identifying the heterogeneity in existing efforts to abstract and measure discrete historical events of interest and to provide theoretical justifications from existing research about what is

[4]Other related datasets that insufficiently overlap ICBe's domain for comparison include BCOW (Leng and Singer 1988), WEIS (McClelland 1978), CREON (Hermann 1984), CASCON (Bloomfield and Moulton 1989), SHERFACS (Sherman 2000), Real-Time Phoenix (Brandt et al. 2018), and COfEE (Balali, Asadpour, and Jafari 2021) (see histories in Merritt (1994) and Schrodt and Hall (2006)).

included in our dataset's ontology and where ICBe's detail about historical events can be compared to the current state of the art.

Table 1: Ontological coverage of ICBe versus existing State of the Art

	Concept	Literature	ICBe	Phoenix	Terrier	ICEWs	MID Incidents	UCDP-GED	ACLED	ICB	MIDs	COW
Domain	Type (Episode or Event)		Ev	Ev	Ev	Ev	Ev	Ev	Ev	Ep	Ep	Ep
	Start		1918	1945	1977	1995	1993	2015	1997	1918	1816	1816
	End		2017	2019	2018	2020	2010	2022	2023	2017	2014	2007
	N		32K	8.5M	28.4M	17.5M	9.6K	128K	1M	1K	5.9K	1K
	Coders (Hand or Automated)		H	A	A	A	H	H	H	H	H	H
	Corpus		ICB	News	News	News	Mix	News	Mix	Mix	Mix	Mix
	Date source (Event or Article)		E	A	A	A	E	A	E	E	E	E
	Location source (Event or Actor)		E	E	E	E	A	E	E	A	E	A
Players	States	Fazal (2011), Spruyt (1996)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Subnational Actors	Haffar (2002)	✓	✓	✓	✓		✓	✓			✓
	IGO/NGO	Bush and Hadden (2019)	✓	✓	✓	✓			✓	✓		✓
	Civilians	Ben-Yehuda and MishaliRam (2006)	✓	✓	✓	✓		✓	✓			
Pieces	Fatalities	Lacina (2006), McNabb Cochran and Long (2017)	✓				✓	✓	✓	✓	✓	✓
	Force Size	Carafano (2014), Goertz and Diehl (1986)	✓									
	Force Domain	Gartzke and Lindsay (2019), Lindsay and Gartzke (2020)	✓	✓	✓	✓						
	Geography (location, territorial change)	Carter (2010)	✓						✓			
Think	Alert (Start/End Crisis)	Brecher and Wilkenfeld (1997)	✓							✓		
	Wishes (Desire/Fear)	Goldgeier and Tetlock (2001)	✓							✓		
	Evaluation (Victory/Defeat)	Stein and Russett (1980)	✓							✓		
	Aims (Territory, Policy, Regime, Preemption)	Sullivan (2007)										
	Awareness (Discover, Become Convinced)	Ramsay (2017), Yarhi-Milo (2013)	✓									
Say	React to past event (Praise, Disapprove, Accept, Reject, Accuse)	O'Neill (2018)	✓	✓	✓	✓				✓		
	Request future event (Appeal, Demand)	Zartman and Faure (2005)	✓	✓	✓	✓	✓					
	Predict future event (Promise, Threaten, Express Intent, Offer Without Condition)	Sechser (2011)	✓	✓	✓	✓	✓				✓	
	Predict with condition (Offer, Ultimatum)	R. Powell (2002)	✓									
	Government (Leadership/Institution Change, Coup, Assasination)	Goemans, Gleditsch, and Chiozza (2009), J. M. Powell and Thyne (2011)	✓	✓	✓	✓			✓			
	By Civilians (Protest/Riot/Strike)	Chenoweth, Hendrix, and Hunter (2019)	✓	✓	✓	✓			✓			
	Against Civilians (Terrorism, Domestic Rights, Mass Killing, Evacuate)	Eck and Hultman (2007), LaFree and Dugan (2007)	✓	✓	✓	✓			✓			
	Diplomacy (Discussion, Meeting, Mediation, Break off negotiations, Withdraw/Expel Diplomats, Propoganda)	Beardsley (2011)	✓	✓	✓	✓				✓		
	Legal Agreements (Sign Agreement, Settle Dispute, Join War on Behalf of, Ally, Mutual Defense Pact, Open Border, Cede Territory, Allow Inspections, Political Succession, Leave Alliance, Terminate Treaty)	Gibler and Sarkees (2004), Owsiak, Cuttner, and Buck (2018)	✓	✓	✓	✓			✓	✓	✓	
	Violate Agreement (Violate Terms of Agreement)	Leeds (2003)	✓									

Table 1: Ontological coverage of ICBe versus existing State of the Art

	Concept	Literature	ICBe	Phoenix	Terrier	ICEWs	MID Incidents	UCDP-GED	ACLED	ICB	MIDs	COW
Do - Unarmed	Mutual Cooperation (Economic cooperation or Aid, Military Cooperation, Intelligence Cooperation, Unspecified)	Leeds (1999)	✓	✓	✓	✓						
	Directed Aid (General Political Support, Economic Aid, Humanitarian Aid, Military Aid, Intelligence Aid, Unspecified Aid)	Yarhi-Milo, Lanoszka, and Cooper (2016)	✓	✓	✓	✓				✓		
Do - Armed	Preparation (Alert, Mobilization, Fortify, Exercise, Weapons Test)	Lai (2004)	✓	✓	✓	✓	✓		✓		✓	
	Maneuver (Deployment, Show of Force, Blockade, No Fly Zone, Border Violation)	Allen, Flynn, and Martinez Machain (2021)	✓	✓	✓	✓	✓		✓		✓	
	Combat (Battle/Clash, Attack, Invasion/Occupation, Bombard, Cease Fire, Retreat)	Fortna (2018), Min (2021)	✓	✓	✓	✓	✓		✓	✓	✓	✓
	Strategic (Declare War, Join War, Continue Fighting, Surrender, End War, Withdraw from War, Switch Sides)	Sarkees and Wayman (2010), Reiter, Stam, and Horowitz (2016)	✓	✓	✓	✓	✓				✓	
	Autonomy (Assert Political Control Over, Assert Autonomy Against, Annex, Reduce Control Over, Decolonize)	Frederick, Hensel, and Macaulay (2017)	✓	✓	✓	✓	✓		✓			

With the exception of large-scale CAMEO dictionary-based systems (the first grouping of columns), our ontology improves upon the the existing state of the art quantitative datasets that ignore important information content about international interactions.[5] We highlight two particular innovations. First, we separate the 'chess pieces' from the 'chess players' in distinguishing between different actors within a state. By virtue of our ontology coding military versus civilian actors and national leaders versus bureaucrats, our data can be used to explore important questions concerning civilian-military relations (Narang and Talmadge 2017), Track Two diplomacy and the role of sub-national actors (Hsu et al. 2020), and the evolution of what actors are engaged in crises - a topic of increasing interest as states engage in gray zone conflict by employing the coast guard or paramilitary mercenaries instead of internationally recognized state militaries (Gannon et al. 2022). Second, we add information about the domains in which actors behave - whether in land, air, sea, space, or cyber - since they differ in their technology, tactics, geography, and purpose (Gartzke and Lindsay 2019). Doing so allows researchers to identify and explain patterns in escalation conditional on the military means states use in conflict. Recent concerns about cross-domain conflict and the effect of new domains of conflict like space and cyber have made this an endeavor of increased interest to practitioners (Gannon 2022).

2 Methodology and data

2.1 Corpus

For our corpus, we select a set of unusually high-quality historical narratives from the International Crisis Behavior (ICB) project ($n = 471$) with coverage spanning 1918-2017 (Supplementary Information (SI) Appendix 1.1)(Brecher et al. 2021; Brecher and Wilkenfeld 1997). ICB defines a crisis as meeting three conditions: (1) an actor perceives a threat to one of more of its basic values, (2) the actor has a finite time horizon for responding to the perceived threat, and (3) the probability of military hostility has increased (Brecher and Wilkenfeld 1982). Crises are a significant focus of detailed single case studies and case comparisons because they provide an opportunity to examine behaviors in international relations short of, or at least prior to, full conflict (Holsti 1965; Paige 1968; Allison and Zelikow 1971; Gavin 2014; Brecher and Wilkenfeld 1982; Iakhnis and James 2019). The corpus is also unique in being designed to be used in a downstream quantitative coding project, meaning each narrative was written by consensus by a small number of scholars using a uniform coding scheme where things like word choice, writing style, and level of specificity were deliberately done in a consistent manner (Hewitt 2001). Case selection was exhaustive based on a survey of world news archives and region experts, cross-checked against other databases of war and conflict, and non-English sources (Kang and Lin 2019; Brecher et al. 2021, 59).

[5] See Balali, Asadpour, and Jafari (2021) for a recent review of ontological depth and availability of Gold Standard example text.

2.2 Coding Process

The ICBe ontology follows a hierarchical design philosophy where a smaller number of significant decisions are made early on and then progressively refined into more specific details (Brust and Denzler 2020).[6] Each coder was instructed to first thoroughly read the full crisis narrative and then presented with a custom graphical user interface (GUI) (SI Appendix 2.1). Coders then proceeded sentence by sentence, choosing the number of events (0-3) that occurred, the highest behavior (thought, speech, or action), a set of players, whether the means were primarily armed or unarmed, whether there was an increase or decrease in aggression (uncooperative/escalating or cooperative/de-escalating), and finally one or more specific and non-mutually exclusive activities. Some additional details were always collected (e.g. location and timing) while other details were only collected if appropriate (e.g. force size, fatalities, domains, units). While each event was matched to a sentence, coders could fill in details outside that sentence (e.g. antecedents to pronouns). We reviewed, standardized, and normalized where coders listed a behavior, actor, or location outside the ontology.[7]

A unique feature of the ontology is that thought, speech, and do behaviors can be nested into combinations, e.g. an offer for the U.S.S.R. to remove missiles from Cuba in exchange for the U.S. removing missiles from Turkey. Through compounding, the ontology can capture what players were said to have known, learned, or said about other specific fully described actions. Two other novel features of our data – its re-programmamable coding and its inclusion of thought, speech, and action – intersect to provide a unique contribution.

First, re-programmable coding means scholars can make modifications to the ontology we provide and re-run the code script to produce a modified dataset. For example, if thoughts and speech acts are immaterial for a particular research question, those higher-level categories can be modified in a way that does not produce inevitable downstream effects on the rest of the coding. This is necessary because unlike existing datasets where variable values are largely independent, the nested nature of the ICBe variables means filtering out speech acts also requires filtered out variables about the actors, locations, and timing of those speech acts.

Second, no existing event data distinguishes thoughts, speeches, and actions. In fact, most only try to code actions and entirely omit thoughts and speech acts despite recognition of their importance in international politics (Smith 1998). Scholars have opted against coding thoughts and speech acts because of a lack of confidence the full universe could be readily observed and consequently at least theoretically be included.[8] But the perfect should not be the enemy of the good, and measurement challenges are only overcome after an initial attempt to estimate difficult to observe concepts of interest. The ICB narratives are one of the better sources for this endeavor due to the consistent use of high-quality primary source material that takes

[6]This process quickly focuses the coder on a smaller number of relevant options while also allowing them to apply multiple tags if the sentence explicitly includes more than one or there is insufficient evidence to choose only one tag. The guided coding process also allows for the possibility that earlier coarse decisions have less error than later fine-grained decisions.

[7]See the full codebook on Github Repository ICBEventData.

[8]Even the coding of overt actions like MIDs is not without contention (Gibler 2018).

advantage of qualitative methods well-suited to identifying thoughts and speech acts like archival work and expert interviews. Important questions about potential bias remain like whether a thought was known at the time by those involved or only recorded long afterwards by historians, but data validation required there to be data to be validated. By design, existing datasets largely omit thoughts and speech acts and also treat the resulting dataset as the static product of hardwired coding decisions. In allowing scholars to identify actions, thoughts, and speech acts as distinct characteristics of international crises while also granting flexibility about the inclusion of a non-random sample of novel variables, our approach is a notable improvement over the state of the art.

Each crisis was typically assigned to 2 expert coders and 2 novice coders with an additional tie-breaking expert coder assigned to sentences with high disagreement.[9] For the purposes of measuring intercoder agreement and consensus, we temporarily disaggregate the unit of analysis to the Coder-Crisis-Sentence-Tag (n=993,731), where a tag is any unique piece of information a coder can associate with a sentence such as an actor, date, behavior, etc. We then aggregate those tags into final events (n=18,783), using a consensus procedure (SI Appendix 2.2) that requires a tag to have been chosen by at least one expert coder and either a majority of expert or novice coders. This screens noisy tags that no expert considered possible but leverages novice knowledge to tie-break between equally plausible tags chosen by experts. Requiring sentence-tag matching may underestimate agreement but minimizes the inclusion of noise and allows for additional validation. Once filtered for agreement, we find 472 actors and 119 different behaviors: 12 thought, 13 speech, and 94 actions.

3 Performance comparison

3.1 Internal consistency

We evaluate the internal validity of the coding process in several ways. For every tag applied we calculate the observed intercoder agreement as the percent of other coders who also applied that same tag (SI Appendix 2.3). Across all concepts, the Top 1 Tag Agreement was low among novices (31%), moderate for experts (65%), and high (73%) following the consensus screening procedure.

We attribute the remaining disagreement primarily to three sources. First, we required coders to rate and justify their confidence in the coding. They reported low confidence for 20% of sentences; 45% of those were due to a mismatch between the ontology and the text ("survey doesn't fit event") and 46% were from a lack of information or confused writing in the source text (40% "more knowledge needed", 6% "confusing sentence"). Observed disagreement varied predictably with self-reported confidence (SI Appendix 2.4). Second, as intended, agreement is higher (75-80%) for questions with fewer options near the root of the ontology compared to agreement for questions near the leaves of the ontology (50%-60%). Third, individual coders

[9]Expert coders were graduate students or postgraduates who collaboratively developed the ontology and documentation for the codebook. Undergraduate coders were students who engaged in classroom workshops.

exhibit nontrivial coding styles, e.g. some more expressive coders applied many tags per concept while others focused on only the single best match. We further observed unintended synonymity, e.g. the same information can be framed as either a threat to do something or a promise not to do something.

3.2 Improvement over existing efforts

To evaluate our coding process relative to existing datasets, we measure the recall and precision of ICBe events in absolute terms and relative to other existing systems. Recall measures the share of desired information recovered by a sequence of coded events while precision measures the degree to which a sequence of events correctly and usefully describes the information in history. To aid in subjective evaluation of the precision and recall of ICBe for each event, we provide full ICB narratives, ICBe coding in an easy-to-read iconographic form, and a wide range of visualizations for every case on the companion website.

Recall for historical episodes is poorly defined for two reasons. History may or may not be written by the victors but by virtue of being written by *someone* there is no genuine ground truth about what occurred, only surviving texts about it (Turberville 1933). Second, there is no *a priori* guide to what information is necessary detail and what is ignorable trivia. History suffers from what is known as the Coastline Paradox (Mandelbrot 1983) — it has a fractal dimension greater than one such that the more you zoom in, the more detail you will find about individual events as well as in between any two discrete events. The ICBe ontology is a proposal about what information is important, but we need an independent benchmark to evaluate whether that proposal is a good one and that allows for comparing proposals from event projects that had different goals. We need a yardstick for history.

Our strategy for dealing with both problems is a plausibly objective yardstick called a synthetic historical narrative. We collect a large diverse corpus of narratives spanning timelines, encyclopedia entries, journal articles, news reports, websites, and government documents. Using natural language processing (fully described in SI Appendix 3.1), we identify details that appear across multiple accounts. A detail refers to the smallest textual unit (a word or n-gram) for which we can calculate similarity across corpora to identify whether sentences semantically refer to the same broader observed event (Narayan, Cohen, and Lapata 2018). The more accounts that mention a detail, the more central it is to understanding the true historical episode. The theoretical motivation is that authors face word limits which force them to pick and choose which details to include, and they choose details that serve the specific context of the document they are producing. With a sufficiently large and diverse corpus of documents, we can vary the context while holding the overall episode constant and see which details tend to be invariant to context. Sufficiently similar details were binned together and then summarized so they could be compared to the coding in ICBe. This presents a harder evaluation baseline than comparing ICBe's recall to just that of ICB since there are non-crisis aspects of these events that may be included in other narratives but are out of the scope of our data. For example, the nationalization of businesses in Cuba may be included as important context in the Cuban Missile Crisis

in documents that do not focus on the crisis dimensions like ICB. Using this hard case, a recall measure of ICBe on the synthetic narratives thus serves as a way to evaluate the breadth of ICBe's ontology and potential application to non-crisis international events.

We find substantive variation in recall across existing state of the art methods. Mentions of a detail across accounts are exponentially distributed with context-invariant details appearing dozens to hundreds of times more than context-dependent details.[10] Furthermore, crisis start and stop dates are arbitrary, and the historical record points to many precursor events as necessary detail for understanding later events. Figure 2 compares ICBe's recall with that of existing datasets for the two case studies detailed in Section 4. ICBe strictly dominates all of the systems but ICEWs in recall though we note that the small sample sizes mean these systems should be considered statistically indistinguishable. Across all existing datasets and ICBe, recall increases with the number of document mentions which is an important sign of validity for both them and our benchmark. The one outlier is Phoenix which in the Cuban Missile Crisis case is so noisy that it's recall curve is flat to decreasing as mentions increase. The two episode-level datasets (MIDs and ICM) have low coverage of contextual details. The two other dictionary systems ICEWs and Terrier have higher coverage, with ICEWs outperforming Terrier. Importantly our corpus of ICB narratives has high recall of frequently mentioned details giving us confidence in how those summaries were constructed, and ICBe lags only slightly behind showing that it left little additional information on the table.[11]

The second component of event measurement validation is precision. It does little good to recall a historical event but too vaguely (e.g. MIDs describes the Cuban Missile Crisis as a blockade, a show of force, and a stalemate) or with too much error to be useful for downstream applications (e.g. ICEWS records 263 "Detonate Nuclear Weapons" events between 1995-2019). ICBe's ontology and coding system are designed to strike a balance so that the most important information is recovered accurately but also abstracted to a level that is still useful and interpretable.

How does ICBe's precision compare to the existing state of the art? A researcher should be able to lay out the events of a crisis on a timeline and read off the macrostructure of an episode from each individual move. We call this visualization a crisis map, a directed graph intersected with a timeline. A crisis map using ICBe for the Cuban Missile Crisis case study is provided in Figure 2, and crisis maps for the two case studies using existing event datasets can be found in SI Appendix 4.3 and 4.4 and crisis maps for all crises using all datasets can be found on the companion website. The crisis maps reveal the episode-level datasets like

[10] As the ICB narratives are intended to explain conflictual behavior in a political context, many of the missing events concern more economic components of conflict (eg. nationalizing a foreign business). Even when they occur in the context of a crisis, these events largely fall outside the sample of information on which ICBe's ontology is currently trained. Even with this limitation, ICBe is more comprehensive than the existing datasets that do try to code the economic dimensions of these crises. We see expanding the ontology to broader international phenomenon as a promising future implementation of our model.

[11] Although Figure 2 focuses only on two crises, the synthetic narrative approach and recall comparison can, and should, be more broadly applied to all international crises in a way that could reveal systematic blindspots across datasets.

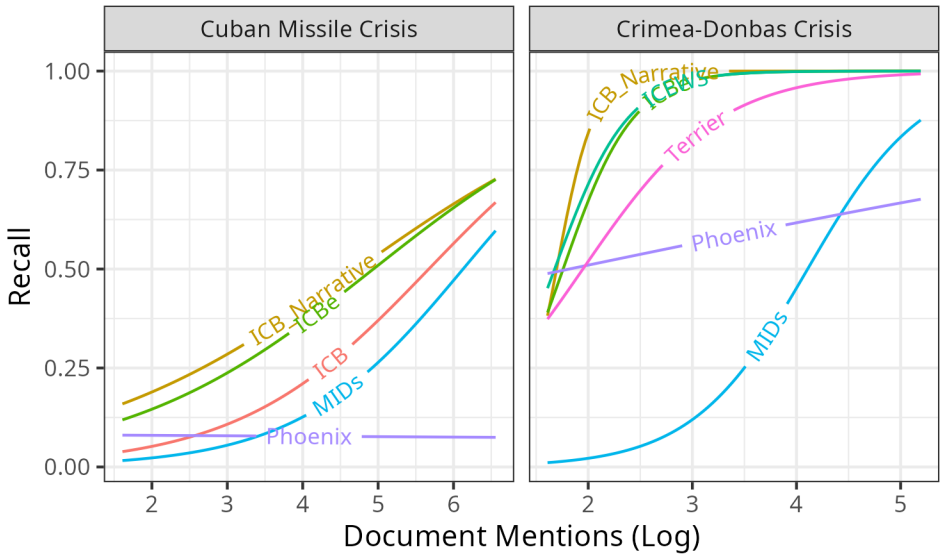

Figure 2: Recall comparison of two cases across existing state of the art efforts. Higher y-axis values represent higher recall and higher x-axis values represent number of times that detail is mentioned across the full corpus used to construct the synthetic narrative.

MIDs or the original ICB are too sparse and vague to reconstruct the structure of the crisis (SI Appendix 4.3 and 4.4). On the other end of the spectrum, the high recall dictionary-based event datasets like Terrier and ICEWs produce so many noisy events (several hundred thousand) that even with heavy filtering their crisis maps are completely unintelligible. Further, because of copyright issues, none of these datasets directly provide the original text spans making event-level precision difficult to verify.

We further want to verify individual event codings, which we can do in the case of ICBe because each event is mapped to a specific span of text. We develop the iconography system for presenting event codings as coherent statements that can be compared side by side to the original source narrative for every case on the companion website. We further provide a stratified sample of event codings alongside their source text (SI Appendix 4.2). We find both the visualizations of macrostructure and head-to-head comparisons of ICBe codings to the raw text to strongly support the quality of ICBe.

Figure 3 shows the location of every sentence from the ICBe corpus in semantic space as embedded using the same large language model as before, and the median location of each ICBe event tag applied to those sentences.[12] Labels reflect the individual leaves of the ontology and colors reflect the higher level coarse branch nodes of the ontology. If ICBe has high precision, substantively similar tags ought to have been applied to substantively similar source text, which is what we see both in two dimensions in the main plot and via hierarchical clustering on all dimensions in the dendrogram along the right-hand side.[13]

[12]We preprocess sentences to replace named entities with a generic Entity token.
[13]Hierarchical clustering on cosine similarity and with Ward's method.

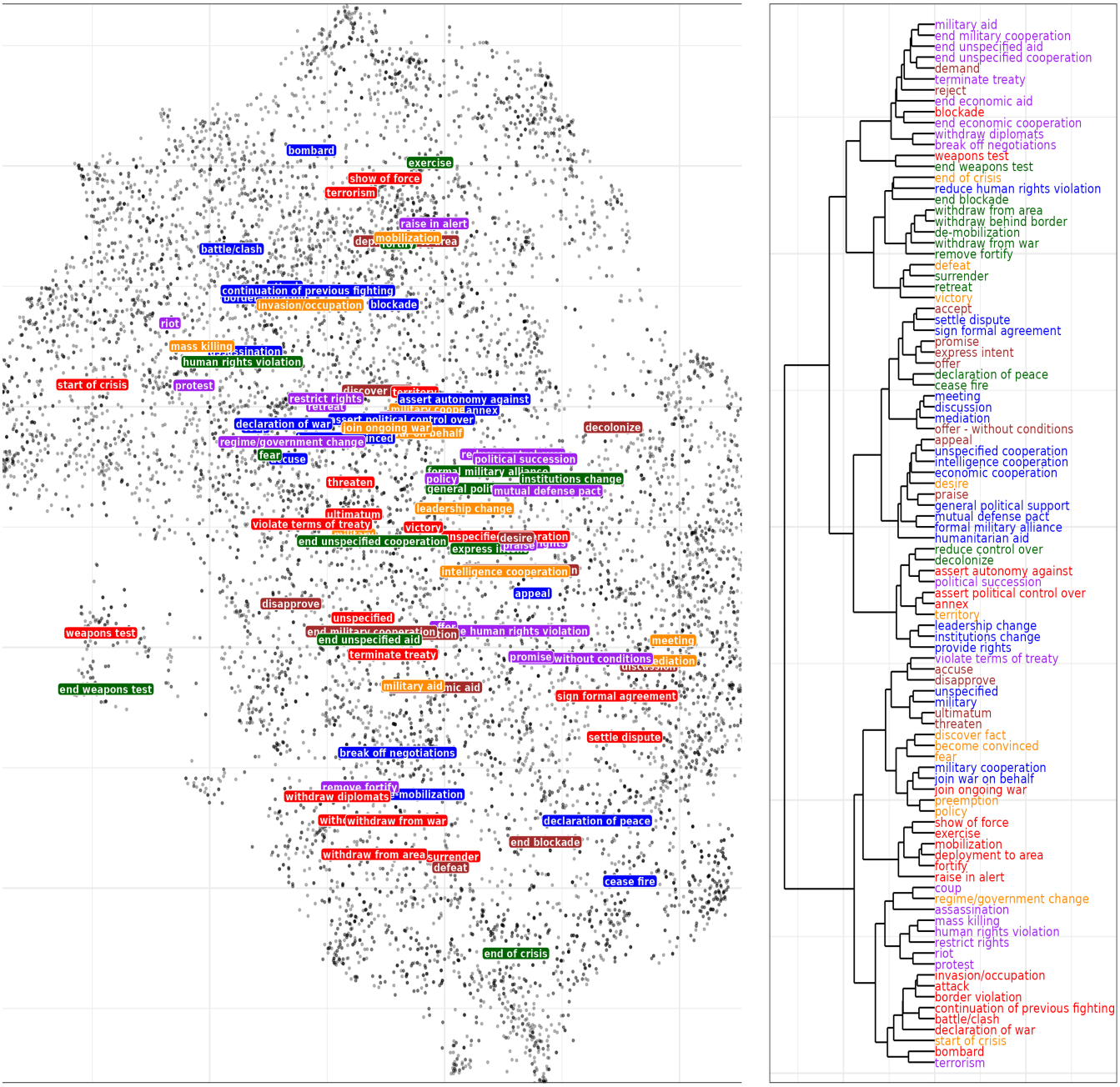

Figure 3: Dots represent individual ICB narrative sentences, as embedded by the Paraphrase-MPNET-base-v2 large language model and flattened into two dimensions with UMAP. Text labels reflect individual leaves of the ICBe ontology, and colors represent intermediate branches of the ontology. Label placement is the median of all of the sentences that the tag was applied to by the coders. The dendrogram shows hierarchical clustering of the tags. If ICBe precision is high, then the sentences that tags were applied to ought to say similar things and the intended shape of the ontology ought to be visually recognizable.

4 Case illustrations

In this section, we focus our validation on two case studies for which we have produced synthetic narratives using the method described in Section 3.2. Our proposed measure is a reconstruction task to see whether our intended ontology can be recovered through only unsupervised clustering of sentences they were applied to. The first (Figure 1) is the Cuban Missile Crisis (hereafter Cuban Missiles) which took place primarily in the second half of 1962, involved the United States, the Soviet Union, and Cuba, and is widely known for bringing the world to the brink of nuclear war. The second (SI Appendix 4.1) is the Crimea-Donbas Crisis (hereafter Crimea-Donbas) which took place primarily in 2014, involved Russia, Ukraine, and NATO, and within a decade spiraled into a full-scale invasion. We choose these cases because they are significant in contemporary international relations, are widely known across academic disciplines as well as among the public, and are sufficiently brief to evaluate in depth. They are similar in that both cases involve a superpower in crisis with a neighbor that changed from a friendly to a hostile regime, both held implications for the economic and military security for the superpower by risking full-scale invasion, and both eventually invited intervention by an opposing superpower.

4.1 Cuban Missile Crisis (1962)

A synthetic historical narrative for Cuban Missiles appears in Figure 4, with 51 events drawn from 2,020 documents. Each row represents a detail that appeared in at least five documents along with an approximate start date, a handwritten summary, the number of documents it was mentioned in, and whether it could be identified in the text of the original ICB corpus, our ICBe events, and any of the competing existing models.

ICB's improved recall of Cuban Missiles was relative to the state of the art was summarized in Section 3.2 (Figure 2), but the events that explain that improvement can now be seen. Our ground truth ICB narrative contains 17/51 of the events from the synthetic narrative of a case that includes high-level previously classified details. ICBe captures nearly all details included in ICB as well as more details from the synthetic narrative than any competing dataset. Phoenix includes some earlier information than ICBe like the nationalization of businesses and back channel negotiations, but the crisis narrative has a clean canonical end with the Soviets agreeing to withdraw missiles. ICBe stands out in including more communicative behavior (do – speech) than existing datasets like US threats to attack and later promises not to invade. Given the recognized importance of threat credibility for understanding international conflict, the addition of this information is a substantively important improvement over the existing state of the art (Slantchev 2011).

Figure 5 shows the crisis map for the Cuban Missile Crisis. Looking at the crisis on a timeline, one can now identify the structure of actors and the environment, along with its supporting details, in a way that validates the precision of ICBe. Although harder to measure objectively, this crisis map provides face valid evidence that ICBe's account is not too vague, but also not unnecessarily detailed. We include much of

YMD	Ground Truth Events	Docs	ICB Corpus	ICBe	ICB	MIDs	Phoenix
1958-12-31	communist coup	128					
1959-06-08	nationalizes owned businesses	28					✓
10-26	backchannel negotiates with	18					✓
1960-01-01	begins recon flights over	10					
03-17	prepares for invasion of	159	✓	✓			
05-01	U-2 downed over	8				✓	
07	establishes diplomatic and trade relations with	10					
07-08	embargos	21					
09-14	attempts assasination	9					
1961-01-03	breaks diplomatic ties	40					
04-17	attempts coup in	16					
	invades	192	✓	✓	✓	✓	
19	Invasion fails	5	✓	✓	✓		
06-01	provides economic and military aid to	43	✓	✓			
08-13	begins construction of Berlin Wall	62	✓	✓	✓	✓	
11-30	covert destabalization efforts against	44					
1962-04-01	places nuclear missiles in	93					
05-21	begins placing nuclear missiles in	31					
07-01	asks for weapons	5					
	deploy troops to	5					
	places nuclear missiles in as response to placing nuclear missiles in	347					
08-10	begins to suspect nuclear missiles will be placed in	29	✓				
09-04	meets to denies presence of missiles	14	✓				
	demands withdrawl and threatens nuclear response on if attacked from	125					
11	threatens war with if attack on or ships	8					
10-01	deploy nuclear armed submarines	5					
14	discovers nuclear missiles in starts crisis	705	✓	✓	✓		
17	mobilizes troops for invasion of	13	✓	✓			
18	meets to denies presence of missiles	21					
22	blockades	400	✓	✓	✓	✓	
	demands withdrawl	7					
	raises nuclear alert	5					
	threatens military attack	10					
23	OAS statement of support for	7	✓	✓			
	meets denies offensive intention	8					
24	respects blockade of	23			✓		
	raises nuclear alert	14					
25	confronts at the	30					✓
26	offers remove missiles for no invasion pledge	60	✓	✓	✓		
27	nuclear missiles in operational	5					
	accidentally violates airspace	25					
	interdicts submarine	16					
	offers withdraw of nuclear missiles in for missiles in	115					
	promises to not invade	32	✓	✓			✓
	U-2 shot down over	119	✓	✓		✓	
28	withdraws nuclear missiles in trade for promise to not invade ends crisis	647	✓	✓	✓	✓	
30	refuses observers	7	✓	✓			
11-20	ends blockade	22	✓	✓	✓		
1963-04-01	removes missiles from	5					
08-05	sign Nuclear Test Ban Treaty	94					
30	install hotline	55					

Figure 4: Synthetic narratives combine several thousand accounts of each crisis into a single timeline of events, taking only those mentioned in at least 5 or more documents. Checkmarks represent whether that event could be hand matched to any detail in the ICB corpus, ICBe dataset, or any of the other event datasets (SI Appendix 3.2 and 3.3).

the geopolitically important details like Soviet deployment, US discovery of that deployment, heightened alert levels, a blockade, and negotiations that ended with a formal agreement. At the same time, the crisis map indicates that ICBe does not include unnecessary nuances that preclude useful comparison to other international events.

4.2 Crimea-Donbas (2014)

A synthetic historical narrative for the 2014 Crimea-Donbas crisis (30 events drawn from 971 documents) appears in Figure 6. As in the earlier case, rows represent details that appeared in at least five documents and whether it is identified in ICBe and existing datasets.

Again quantitatively summarized earlier in Section 3.2 (Figure 2), our ground truth ICB narrative contains 23/30 of the events from the synthetic narrative. Like the gray zone precursor to the Cuban Missile crisis

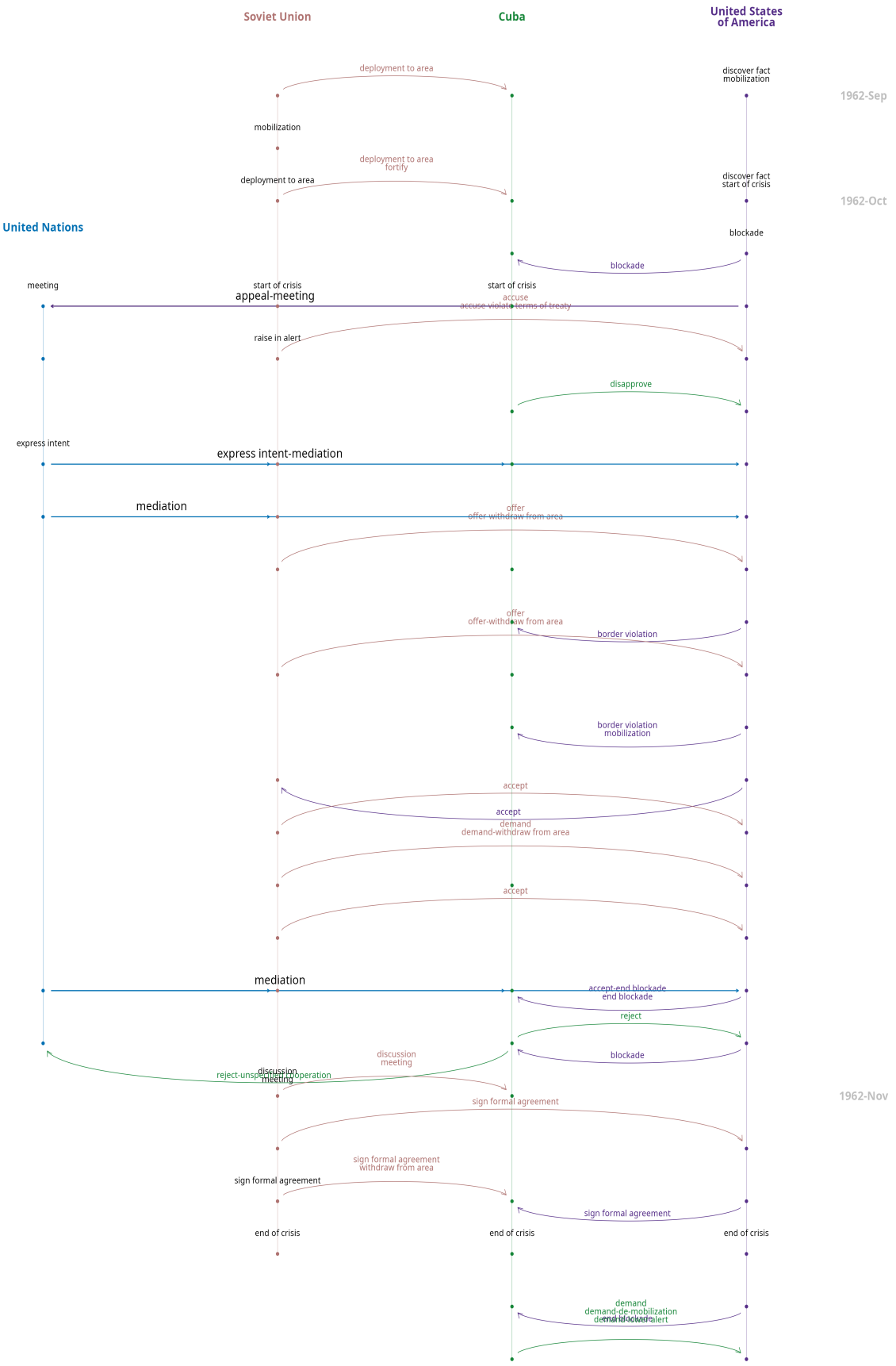

Figure 5: Crisis map for the Cuban Missile Crisis. The start of the crisis is at the top and end of the crisis is the bottom, with each actor in a column with labeled points identifying their speeches, actions, and thoughts.

YMD	Ground Truth Events	Docs	ICB Corpus	ICBe	MIDs	phoenix	terrier	icews
1994-12-05	nuclear disarmed in exchange for promise to never invade	7						
2004-04-02	NATO expands to include Estonia, Latvia, and Lithuania	5						
2008-01-01	NATO provides incredile offer of membership to	7	✓	✓				
08-01	invades Georgia and annexes Abkhazia and Ossetia	10	✓	✓	✓			
2010-04-21	and ratified the Russian Ukrainian Naval Base for Gas treaty, extending the Russian Navy's lease of n facilities for 25 years after 2017	5						
2013-11-01	plans to join trade agreement	6	✓			✓		✓
21	rejects trade agreement	42	✓	✓			✓	✓
	protests	6	✓	✓		✓	✓	✓
12-17	offers debt relief and discounted energy to	5				✓	✓	✓
2014-02-01	provide econ and military aid	26	✓	✓		✓	✓	✓
21	backs political settlement	10	✓	✓		✓	✓	✓
22	leader removed and flees to	65	✓	✓		✓	✓	✓
27	backed gunmen begin seizing government buildings in	10	✓	✓		✓	✓	✓
03-01	votes to deploy military to	11	✓	✓		✓	✓	✓
16	independence refferendum	12	✓					✓
	mobilizes forces to and border	11	✓	✓		✓	✓	✓
	backed seperatists attack in Donbas	6				✓	✓	✓
17	sanctions	83	✓	✓	✓	✓	✓	✓
18	annexes	180	✓	✓	✓		✓	✓
04-06	backed seperatists begin fighting in (Donetsk and Luhansk)	146	✓	✓	✓	✓	✓	✓
23	military excercises at border	7	✓	✓			✓	✓
05-11	DPR and the LPR declare independence	5	✓	✓				
21	DPR requests Russian military intervention	5				✓		
07-17	shoots down passenger jet (Malaysia Airlines Flight 17)	6	✓	✓				✓
	sanctions	5	✓	✓		✓	✓	✓
	econ sanction	5	✓	✓		✓	✓	✓
09-05	, , the DPR and the LPR signed a ceasefire agreement, the Minsk I	10	✓	✓			✓	✓
06	, DPR, and the LPR, continue fight with	15	✓	✓		✓		✓
2015-02-12	, , the DPR and the LPR signed a ceasefire agreement (Minsk II)	10	✓	✓			✓	✓
21	recognizes Donetsk People's Republic and the Luhansk People's Republic	6						

Figure 6: Synthetic narratives combine several thousand accounts of each crisis into a single timeline of events, taking only those mentioned in at least 5 or more documents. Checkmarks represent whether that event could be hand matched to any detail in the ICB corpus, ICBe dataset, or any of the other event datasets (SI Appendix 3.2 and 3.3).

(Cormac and Aldrich 2018), Ukraine provided several security guarantees to Russia that were potentially undone, e.g. a long term lease on naval facilities in Crimea. But unlike the Cuban Missile crisis, the end of this crisis is unclear, with the event meekly ending with a second cease-fire agreement (Minsk II) but continued fighting. ICBe again recalls more important information about the crisis than any existing dataset, particularly information concerning the behavior of non-state separatist groups like the Donetsk People's Republic (DPR) and Luhansk People's Republic (LPR).

As this more recent case reflects primarily public reporting rather than the previously classified details relevant for the Cuban Missile Crisis, ICBe's improvement relative to the global and real-time coverage of dictionary-based event systems is still present, but less pronounced. We want to take seriously the possibility that some functional transformation could recover the precision of ICBe. For example, Terechshenko (2020) attempts to correct for the mechanically increasing amount of news coverage each year by de-trending violent event counts from Phoenix using a human-coded baseline. Others have focused on verifying precision for ICEWs on specific subsets of details against known ground truths, e.g. geolocation (Cook and Weidmann 2019), protest events (80%) (Wüest and Lorenzini 2020), anti-government protest networks (46.1%) (Jäger 2018).

We take the same approach here in Figure 7, selecting four specific CAMEO event codings and checking how often they reflect a true real-world event from the Crimea-Donbas synthetic narrative. We choose four event types around key moments in the crisis. The start of the crisis revolves around Ukraine backing out of a trade deal with the EU in favor of Russia, but "sign formal agreement" events act more like a topic detector with dozens of events generated by discussions of a possible agreement but not the actual agreement which never materialized. The switch is caught by the "reject plan, agreement to settle dispute", but also continues for Viktor Yanukovych even after he was removed from power because of articles retroactively discussing the cause of his removal. Events for "use conventional military force" capture a threshold around the start of hostilities and who the participants were but not any particular battles or campaigns. Likewise, "impose embargo, boycott, or sanctions" captures the start of waves of sanctions and from who but are effectively constant as the news coverage does not distinguish between subtle changes or additions. In sum, dictionary-based methods on news corpora tend to have high recall because they parse everything in the news, but for the same reason, their specificity for most event types is too low to back out individual chess-like sequencing that ICBe aims to record.

5 Conclusion

We investigated event abstraction from narratives describing key historical episodes in international relations. We synthesized a prior belief about the latent unobserved phenomena that drive these events in international relations and proposed a mapping to observable concepts that enter into the observed historical record. We designed an ontology with high coverage over those concepts and developed a training procedure and technical stack for human coding of historical texts. Multiple validity checks find the resulting codings have high internal validity (e.g. intercoder agreement) and external validity (i.e. matching source material in both micro-details at the sentence level and macro-details spanning full historical episodes). Further, these codings perform much better in terms of recall, precision, coverage, and overall coherence in capturing these historical episodes than existing event systems used in international relations.

We release several open-source products along with supporting code and documentation to further advance the study of international relations, event extraction, and natural language processing. The first is the International Crisis Behavior Events (ICBe) dataset, an event-level aggregation of what took place during the crises identified by the ICB project. These data are appropriate for statistical analysis of hard questions about the sequencing of events (e.g. escalation and de-escalation of conflicts).[14] Second, we provide a coder-level disaggregation with multiple codings of each sentence by experts and undergrads that allows for the introduction of uncertainty and human interpretation of events. Further, we release a direct mapping from the codings to the source text at the sentence level as a new resource for natural language processing.

[14] Using ICBe data, Gannon (2022) finds that cross-domain crises are shorter and less violent than same-domain crises.

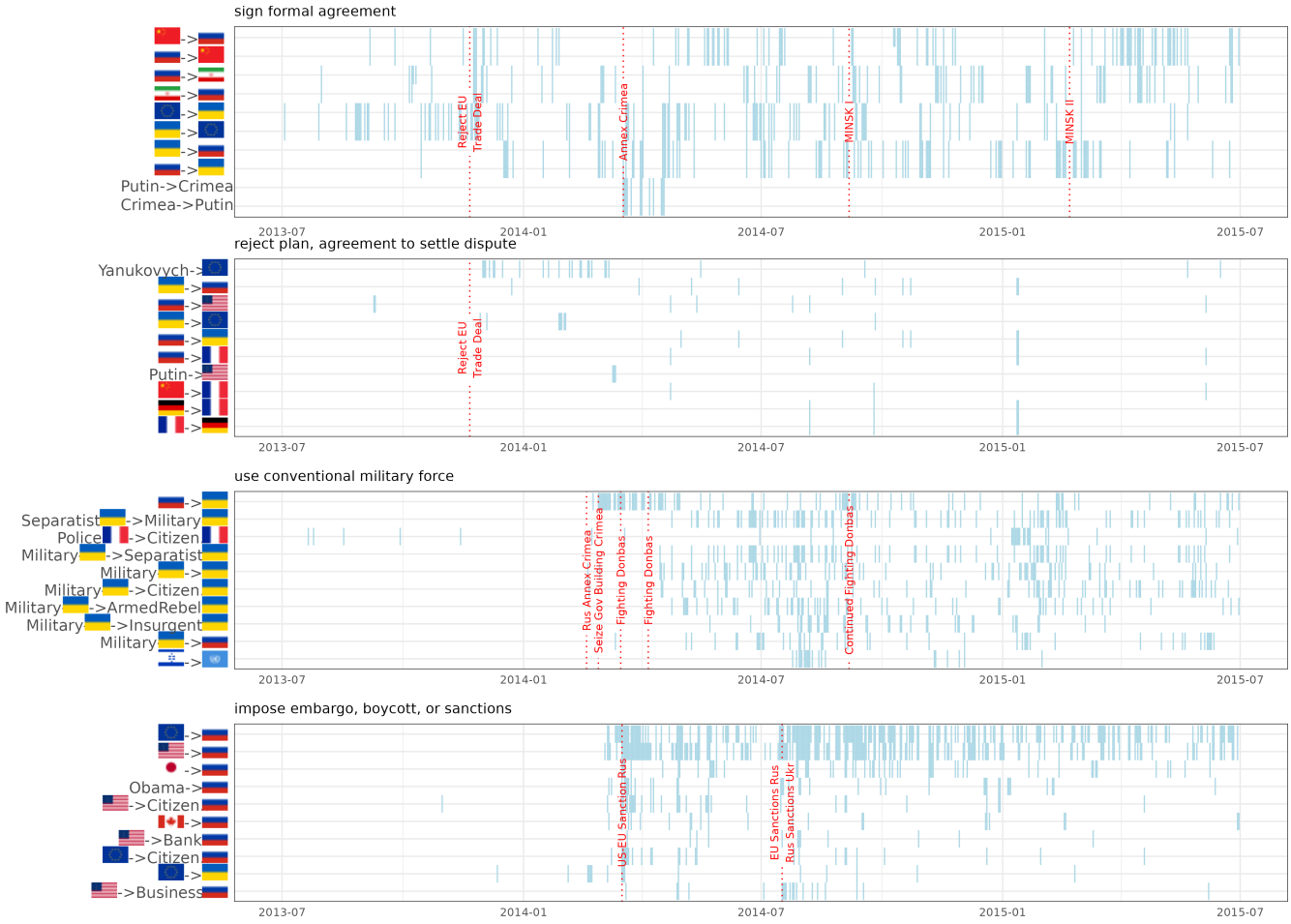

Figure 7: The unit of analysis is the dyad-day. Top 10 most active dyads per category shown. Red text shows events from the synthetic narrative relative to that event category. Blue bars indicate an event recorded by ICEWs for that dyad on that day.

Finally, we provide a companion website that incorporates detailed visualizations of all of the data introduced here at crisisevents.org.

6 Funding

This work was supported by a grant from the Office of Naval Research $N00014-19-1-2491$ and from the Charles Koch Foundation 20180481. The financial sponsors played no in the design, execution, analysis and interpretation of data, or writing of the study.

7 Acknowledgements

We thank the ICB Project and its directors and contributors for their foundational work and their help with this effort. We make special acknowledgement of Michael Brecher for helping found the ICB project in 1975, creating a resource that continues to spark new insights to this day. We thank the many undergraduate

coders. Thanks to the Center for Peace and Security Studies and its membership for comments. Special thanks to Rebecca Cordell, Philip Schrodt, Zachary Steinert-Threlkeld, and Zhanna Terechshenko for their generous feedback. Thank you to the cPASS research assistants: Helen Chung, Daman Heer, Syeda ShahBano Ijaz, Anthony Limon, Erin Ling, Ari Michelson, Prithviraj Pahwa, Gianna Pedro, Tobias Stodiek, Yiyi 'Effie' Sun, Erin Werner, Lisa Yen, and Ruixuan Zhang.

8 Author Contributions

Conceptualization: R.W.D., E.G., J.L.; Methodology: R.W.D., T.L.S.; Software: R.W.D.; Validation: R.W.D., T.L.S.; Formal Analysis: R.W.D., T.L.S.; Investigation: S.C., R.W.D., J.A.G., C.K., N.L., E.M., J.M.C.N., D.P., D.Q., J.W.; Data Curation: R.W.D., D.Q., T.L.S., J.W.; Writing - Original Draft: R.W.D., T.L.S.; Writing - Review & Editing: R.W.D., J.A.G., E.G., T.L.S.; Visualization: R.W.D., T.L.S.; Supervision: E.G.; Project Administration: S.C., R.W.D., J.A.G., D.Q., T.L.S., J.W.; Funding Acquisition: E.G., J.L.

9 Data Availability Statement

This article's data, supplementary appendix, replication material, and visualizations of every historical episode are available on the GitHub repository ICBEventData and through the companion website crisisevents.org.

10 Competing Interests Declaration

The authors declare that there are no competing interests.

Works Cited

Allen, Michael A, Michael E Flynn, and Carla Martinez Machain. 2021. "US Global Military Deployments, 1950–2020*." *Conflict Management and Peace Science*, July, 07388942211030885. https://doi.org/10.1177/07388942211030885.

Allison, Graham T., and Philip Zelikow. 1971. *Essence of Decision: Explaining the Cuban Missile Crisis*. Vol. 327. Little, Brown Boston.

Althaus, Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. 2019. "Cline Center Historical Phoenix Event Data Variable Descriptions." *Cline Center Historical Phoenix Event Data*.

Balali, Ali, Masoud Asadpour, and Seyed Hossein Jafari. 2021. "COfEE: A Comprehensive Ontology for Event Extraction from Text." arXiv. https://doi.org/10.48550/arXiv.2107.10326.

Beardsley, Kyle. 2011. *The Mediation Dilemma*. Cornell University Press.

Beardsley, Kyle, Patrick James, Jonathan Wilkenfeld, and Michael Brecher. 2020. "The International Crisis Behavior Project." *Oxford Research Encyclopedia of Politics*. https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-1638. https://doi.org/10.1093/acrefore/9780190228637.013.1638.

Beger, Andreas, Richard K. Morgan, and Michael D. Ward. 2021. "Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict." *Journal of Conflict Resolution* 65 (7-8): 1405–26. https://doi.org/10.1177/0022002720982358.

Beieler, John, Patrick T. Brandt, Andrew Halterman, Philip A. Schrodt, Erin M. Simpson, and R. Michael Alvarez. 2016. "Generating Political Event Data in Near Real Time." In *Computational Social Science*, 98. Cambridge University Press.

Ben-Yehuda, Hemda, and Meirav MishaliRam. 2006. "Ethnic Actors and International Crises: Theory and Findings, 1918–2001." *International Interactions* 32 (1): 49–78. https://doi.org/10.1080/03050620600584435.

Bloomfield, Lincoln P., and Allen Moulton. 1989. "CASCON III: Computer-aided System for Analysis of Local Conflicts." *MIT Center for International Studies, Cambridge*.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. "ICEWS Coded Event Data." *Harvard Dataverse* 12.

Brandt, Patrick T., Vito DOrazio, Jennifer Holmes, Latifur Khan, and Vincent Ng. 2018. "Phoenix Real-Time Event Data."

Brecher, Michael. 1999. "International Studies in the Twentieth Century and Beyond: Flawed Dichotomies, Synthesis, Cumulation: ISA Presidential Address." *International Studies Quarterly* 43 (2): 213–64.

Brecher, Michael, and Jonathan Wilkenfeld. 1982. "Crises in World Politics." *World Politics* 34 (3): 380–417. https://doi.org/10.2307/2010324.

———. 1997. *A Study of Crisis*. University of Michigan Press.

Brecher, Michael, Jonathan Wilkenfeld, Kyle C. Beardsley, Patrick James, and David Quinn. 2021. "International Crisis Behavior Data Codebook." Codebook Version 14.

Brust, Clemens-Alexander, and Joachim Denzler. 2020. "Integrating Domain Knowledge: Using Hierarchies to Improve Deep Classifiers." *arXiv:1811.07125 [Cs]*, January. https://arxiv.org/abs/1811.07125.

Bush, Sarah Sunn, and Jennifer Hadden. 2019. "Density and Decline in the Founding of International NGOs in the United States." *International Studies Quarterly* 63 (4): 1133–46. https://doi.org/10.1093/isq/sqz061.

Carafano, James Jay. 2014. "Measuring Military Power." *Strategic Studies Quarterly* 8 (3): 11–18. https://www.jstor.org/stable/26270616.

Carter, David B. 2010. "The Strategy of Territorial Conflict." *American Journal of Political Science* 54 (4): 969–87. https://doi.org/10.1111/j.1540-5907.2010.00471.x.

Chenoweth, Erica, Cullen S Hendrix, and Kyleanne Hunter. 2019. "Introducing the Nonviolent Action in Violent Contexts (NVAVC) Dataset." *Journal of Peace Research* 56 (2): 295–305. https://doi.org/10.1177/0022343318804855.

Cook, Scott J., and Nils B. Weidmann. 2019. "Lost in Aggregation: Improving Event Analysis with Report-Level Data." *American Journal of Political Science* 63 (1): 250–64. https://doi.org/10.1111/ajps.12398.

Cormac, Rory, and Richard J. Aldrich. 2018. "Grey Is the New Black: Covert Action and Implausible Deniability." *International Affairs* 94 (3): 477–94. https://doi.org/10.1093/ia/iiy067.

Davies, Shawn, Therése Pettersson, and Magnus Öberg. 2022. "Organized Violence 1989–2021 and Drone Warfare." *Journal of Peace Research* 59 (4): 593–610.

Eck, Kristine, and Lisa Hultman. 2007. "One-Sided Violence Against Civilians in War: Insights from New Fatality Data." *Journal of Peace Research* 44 (2): 233–46. https://doi.org/10.1177/0022343307075124.

Fazal, Tanisha M. 2011. *State Death: The Politics and Geography of Conquest, Occupation, and Annexation. State Death*. Princeton University Press. https://doi.org/10.1515/9781400841448.

Felbermayr, Gabriel, Aleksandra Kirilakha, Constantinos Syropoulos, Erdal Yalcin, and Yoto V. Yotov. 2020. "The Global Sanctions Data Base." *European Economic Review* 129 (October): 103561. https://doi.org/10.1016/j.euroecorev.2020.103561.

Fortna, Virginia Page. 2018. *Peace Time*. Princeton University Press.

Frederick, Bryan A, Paul R Hensel, and Christopher Macaulay. 2017. "The Issue Correlates of War Territorial Claims Data, 1816–20011." *Journal of Peace Research* 54 (1): 99–108. https://doi.org/10.1177/0022343316676311.

Gannon, J Andrés. 2022. "One If by Land, and Two If by Sea: Cross-Domain Contests and the Escalation of International Crises." *International Studies Quarterly* 66 (4): sqac065. https://doi.org/10.1093/isq/sqac065.

Gannon, J Andrés, Erik Gartzke, Jon R. Lindsay, and Peter Schram. 2022. "The Shadow of Deterrence: Why Capable Actors Engage in Contests Short of War." *Journal of Conflict Resolution*, 00220027231166345.

Gartzke, Erik, and Jon R. Lindsay. 2019. *Cross-Domain Deterrence: Strategy in an Era of Complexity*. Oxford University Press.

Gavin, Francis J. 2014. "History, Security Studies, and the July Crisis." *Journal of Strategic Studies* 37 (2): 319–31. https://doi.org/10.1080/01402390.2014.912916.

Gibler, Douglas M. 2018. *International Conflicts, 1816-2010: Militarized Interstate Dispute Narratives*. Rowman & Littlefield.

Gibler, Douglas M., and Meredith Reid Sarkees. 2004. "Measuring Alliances: The Correlates of War Formal Interstate Alliance Dataset, 1816–2000." *Journal of Peace Research* 41 (2): 211–22. https://doi.org/10.1177/0022343304041061.

Glaser, Charles L. 2000. "The Causes and Consequences of Arms Races." *Annual Review of Political Science* 3 (1): 251–76. https://doi.org/10.1146/annurev.polisci.3.1.251.

Goemans, Henk E., Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46 (2): 269–83.

Goertz, Gary, and Paul F. Diehl. 1986. "Measuring Military Allocations: A Comparison of Different Approaches." *Journal of Conflict Resolution* 30 (3): 553–81. https://doi.org/10.1177/0022002786030003009.

Goldgeier, J. M., and P. E. Tetlock. 2001. "Psychology and International Relations Theory." *Annual Review of Political Science* 4 (1): 67–92. https://doi.org/10.1146/annurev.polisci.4.1.67.

Grant, Christan, Andrew Halterman, Jill Irvine, Yan Liang, and Khaled Jabr. 2017. "OU Event Data Project," December.

Haffar, Warren. 2002. "Emergent Peacemakers: Cataloguing New Patterns of Activity in Post-Cold War Conflict." *Peace Economics, Peace Science and Public Policy* 8 (2). https://doi.org/10.2202/1554-8597.1054.

Hermann, Charles. 1984. "Comparative Research on the Events of Nations (CREON) Project: Foreign Policy Events, 1959-1968: Version 1." ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR05205.V1.

Hewitt, J. Joseph. 2001. "Engaging International Data in the Classroom: Using the ICB Interactive Data Library to Teach Conflict and Crisis Analysis." *International Studies Perspectives* 2 (4): 371–83. https://doi.org/10.1111/1528-3577.00066.

Holsti, Ole R. 1965. "The 1914 Case." *The American Political Science Review* 59 (2): 365–78. https://doi.org/10.2307/1953055.

Hsu, Angel, Niklas Höhne, Takeshi Kuramochi, Virginia Vilariño, and Benjamin K. Sovacool. 2020. "Beyond States: Harnessing Sub-National Actors for the Deep Decarbonisation of Cities, Regions, and Businesses."

Energy Research & Social Science 70 (December): 101738. https://doi.org/10.1016/j.erss.2020.101738.

Iakhnis, Evgeniia, and Patrick James. 2019. "Near Crises in World Politics: A New Dataset." *Conflict Management and Peace Science*, July, 0738894219855610. https://doi.org/10.1177/0738894219855610.

Jäger, Kai. 2018. "The Limits of Studying Networks with Event Data: Evidence from the ICEWS Dataset." *Journal of Global Security Studies* 3 (4): 498–511.

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30 (2): 167–214. https://doi.org/10.2307/2009958.

Kang, David C., and Alex Yu-Ting Lin. 2019. "US Bias in the Study of Asian Security: Using Europe to Study Asia." *Journal of Global Security Studies* 4 (3): 393–401.

Kinne, Brandon J. 2020. "The Defense Cooperation Agreement Dataset (DCAD)." *Journal of Conflict Resolution* 64 (4): 729–55. https://doi.org/10.1177/0022002719857796.

Lacina, Bethany. 2006. "Explaining the Severity of Civil Wars." *Journal of Conflict Resolution* 50 (2): 276–89. https://doi.org/10.1177/0022002705284828.

LaFree, Gary, and Laura Dugan. 2007. "Introducing the Global Terrorism Database." *Terrorism and Political Violence* 19 (2): 181–204.

Lai, Brian. 2004. "The Effects of Different Types of Military Mobilization on the Outcome of International Crises." *Journal of Conflict Resolution* 48 (2): 211–29.

Leeds, Brett Ashley. 1999. "Domestic Political Institutions, Credible Commitments, and International Cooperation." *American Journal of Political Science* 43 (4): 979–1002. https://doi.org/10.2307/2991814.

———. 2003. "Alliance Reliability in Times of War: Explaining State Decisions to Violate Treaties." *International Organization* 57 (4): 801–27. https://doi.org/10.1017/S0020818303574057.

Leng, Russell J., and J. David Singer. 1988. "Militarized Interstate Crises: The BCOW Typology and Its Applications." *International Studies Quarterly* 32 (2): 155–73. https://doi.org/10.2307/2600625.

Li, Qian, Hao Peng, Jianxin Li, Yiming Hei, Rui Sun, Jiawei Sheng, Shu Guo, et al. 2021. "A Comprehensive Survey on Schema-based Event Extraction with Deep Learning." *arXiv:2107.02126 [Cs]*, August. https://arxiv.org/abs/2107.02126.

Lindsay, Jon R., and Erik Gartzke. 2020. "Politics by Many Other Means: The Comparative Strategic Advantages of Operational Domains." *Journal of Strategic Studies* 0 (0): 1–34. https://doi.org/10.1080/01402390.2020.1768372.

Lupton, Danielle L. 2018. "Reexamining Reputation for Resolve: Leaders, States, and the Onset of International Crises." *Journal of Global Security Studies* 3 (2): 198–216. https://doi.org/10.1093/jogss/ogy004.

Mandelbrot, Benoit B. 1983. *The fractal geometry of nature*. New York: Freeman.

McClelland, Charles. 1978. "World Event/Interaction Survey, 1966-1978." *WEIS Codebook ICPSR* 5211.

McNabb Cochran, Kathryn, and Stephen B. Long. 2017. "Measuring Military Effectiveness: Calculating Casualty Loss-Exchange Ratios for Multilateral Wars, 1816–1990." *International Interactions* 43 (6): 1019–40. https://doi.org/10.1080/03050629.2017.1273914.

Merritt, Richard L. 1994. "Measuring Events for International Political Analysis." *International Interactions* 20 (1-2): 3–33.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41. https://doi.org/10.1145/219717.219748.

Min, Eric. 2021. "Interstate War Battle Dataset (1823–2003)." *Journal of Peace Research* 58 (2): 294–303. https://doi.org/10.1177/0022343320913305.

Moyer, Jonathan D, Sara D Turner, and Collin J Meisel. 2020. "What Are the Drivers of Diplomacy? Introducing and Testing New Annual Dyadic Data Measuring Diplomatic Exchange." *Journal of Peace Research*, September, 0022343320929740. https://doi.org/10.1177/0022343320929740.

Narang, Vipin, and Caitlin Talmadge. 2017. "Civil-Military Pathologies and Defeat in War: Tests Using New Data." *Journal of Conflict Resolution*, January, 0022002716684627. https://doi.org/10.1177/0022002716684627.

Narayan, Shashi, Shay B Cohen, and Mirella Lapata. 2018. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." *arXiv Preprint arXiv:1808.08745*.

O'Neill, Barry. 2018. "International Negotiation: Some Conceptual Developments." *Annual Review of Political Science* 21 (1): 515–33. https://doi.org/10.1146/annurev-polisci-031416-092909.

Owsiak, Andrew P, Allison K Cuttner, and Brent Buck. 2018. "The International Border Agreements Dataset." *Conflict Management and Peace Science* 35 (5): 559–76. https://doi.org/10.1177/0738894216646978.

Paige, Glenn D. 1968. *The Korean Decision, June 24-30, 1950*. Free Press.

Palmer, Glenn, Roseanne W McManus, Vito D'Orazio, Michael R Kenwick, Mikaela Karstens, Chase Bloch, Nick Dietrich, Kayla Kahn, Kellan Ritter, and Michael J Soules. 2021. "The MID5 Dataset, 2011–2014: Procedures, Coding Rules, and Description." *Conflict Management and Peace Science*, February, 0738894221995743. https://doi.org/10.1177/0738894221995743.

Powell, Jonathan M, and Clayton L Thyne. 2011. "Global Instances of Coups from 1950 to 2010: A New Dataset." *Journal of Peace Research* 48 (2): 249–59. https://doi.org/10.1177/0022343310397436.

Powell, Robert. 2002. "Bargaining Theory and International Conflict." *Annual Review of Political Science* 5 (1): 1–30. https://doi.org/10.1146/annurev.polisci.5.092601.141138.

Quinn, David, Jonathan Wilkenfeld, Kathleen Smarick, and Victor Asal. 2006. "Power Play: Mediation in Symmetric and Asymmetric International Crises." *International Interactions* 32 (4): 441–70. https://doi.org/10.1080/03050620601011107.

Raleigh, Clionadh, Andrew Linke, Håavard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47 (5): 651–60.

Ralph Sundberg, and Mihai Croicu. 2016. "UCDP GED Codebook Version 5.0." Department of Peace and Conflict Research, Uppsala University.

Ramsay, Kristopher W. 2017. "Information, Uncertainty, and War." *Annual Review of Political Science* 20 (1): 505–27. https://doi.org/10.1146/annurev-polisci-051215-022729.

Reiter, Dan. 2015. "Should We Leave Behind the Subfield of International Relations?" *Annual Review of Political Science* 18 (1): 481–99. https://doi.org/10.1146/annurev-polisci-053013-041156.

Reiter, Dan, Allan C. Stam, and Michael C. Horowitz. 2016. "A Revised Look at Interstate Wars, 1816–2007." *Journal of Conflict Resolution* 60 (5): 956–76. https://doi.org/10.1177/0022002714553107.

Sarkees, Meredith Reid, and Frank Wayman. 2010. *Resort to War: 1816-2007*. CQ Press.

Schrodt, Philip A., and Blake Hall. 2006. "Twenty Years of the Kansas Event Data System Project." *The Political Methodologist* 14 (1): 2–8.

Sechser, Todd S. 2011. "Militarized Compellent Threats, 1918–2001." *Conflict Management and Peace Science* 28 (4): 377–401. https://doi.org/10.1177/0738894211413066.

Sherman, Frank L. 2000. "SHERFACS: A Cross-Paradigm, Hierarchical, and Contextually-Sensitive International Conflict Dataset, 1937-1985: Version 1." ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR02292.V1.

Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. Cambridge University Press.

Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92 (3): 623–38. https://doi.org/10.2307/2585485.

Spruyt, Hendrik. 1996. *The Sovereign State and Its Competitors: An Analysis of Systems Change*. Princeton University Press.

Stein, Arthur A., and Bruce M. Russett. 1980. "Evaluating War: Outcomes and Consequences." In *Handbook of Political Conflict: Theory and Research*, 399–422. Free Press New York.

Steinert-Threlkeld, Zachary C. 2019. "The Future of Event Data Is Images." *Sociological Methodology* 49 (1): 68–75. https://doi.org/10.1177/0081175019860238.

Sullivan, Patricia L. 2007. "War Aims and War Outcomes: Why Powerful States Lose Limited Wars." *Journal of Conflict Resolution* 51 (3): 496–524. https://doi.org/10.1177/0022002707300187.

Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–32.

Terechshenko, Zhanna. 2020. "Hot Under the Collar: A Latent Measure of Interstate Hostility." *Journal of Peace Research* 57 (6): 764–76. https://doi.org/10.1177/0022343320962546.

Trager, Robert F. 2016. "The Diplomacy of War and Peace." *Annual Review of Political Science* 19 (1): 205–28. https://doi.org/10.1146/annurev-polisci-051214-100534.

Turberville, A. S. 1933. "History Objective and Subjective." *History* 17 (68): 289–302. https://www.jstor.org/stable/24400365.

Ward, Michael D., Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. 2013. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction." *International Studies Review* 15 (4): 473–90.

Wilkenfeld, Jonathan, and Michael Brecher. 2000. "Interstate Crises and Violence: Twentieth-Century Findings." *Handbook of War Studies II*, 282–300.

Wüest, Bruno, and Jasmine Lorenzini. 2020. "External Validation of Protest Event Analysis." *Contention in Times of Crisis: Recession and Political Protest in Thirty Euro-Pean Countries*, 49–78.

Yarhi-Milo, Keren. 2013. "In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries." *International Security* 38 (1): 7–51. https://doi.org/10.1162/ISEC/_a/_00128.

Yarhi-Milo, Keren, Alexander Lanoszka, and Zack Cooper. 2016. "To Arm or to Ally? The Patron's Dilemma and the Strategic Logic of Arms Transfers and Alliances." *International Security* 41 (2): 90–139. https://doi.org/10.1162/ISEC/_a/_00250.

Zartman, I. William, and Guy Olivier Faure. 2005. *Escalation and Negotiation in International Conflicts*. Cambridge University Press.

Zhang, Han, and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49 (1): 1–57. https://doi.org/10.1177/0081175019860244.