Automated Nuclear Command and Control:

Simulating Two-Way Interactions

Kayla Manning

December 8, 2021

1 Introduction

1.1 Background

Computers can distill large volumes of information into seemingly objective output, unswayed by the emotional factors that affect humans. Given the propensity of computer models to streamline complex and high-stakes decisions, algorithms have played an increasingly important role in decision-making over the years. With applications ranging from criminal justice to healthcare,[1] these decision-making tools have natural extensions to problems in national security. After all, the United States has already incorporated artificial intelligence into military operations in Iraq and Syria, and the government continues to research AI applications in command and control.[2]

While the use of artificial intelligence in national security marks a more recent innovation, the Soviet Union took steps to automate its command and control apparatus through the late Cold War years. Fearing the relative superiority of the United States, the Soviets built a computer program to constantly assess the likelihood of a surprise nuclear attack from the US. If the model indicated an attack was imminent, leadership could prepare to react with a preemptive strike or some other form of precaution.[3]

1. Kosuke Imai and Zhichao Jiang, "Principal Fairness for Human and Algorithmic Decision-Making," January 2021, accessed November 28, 2021, https://imai.fas.harvard.edu/research/fairness.html; Thomas Grote and Philipp Berens, "On the ethics of algorithmic decision-making in healthcare" [in eng], *Journal of Medical Ethics* 46, no. 3 (March 2020): 205–211, ISSN: 1473-4257, https://doi.org/10.1136/medethics-2019-105586.

2. Kelley M Sayler and Daniel S Hoadley, "Artificial Intelligence and National Security" [in en], November 2020, 43.

3. President's Foreign Intelligence Advisory Board, *The Soviet "War Scare"*, Intelligence Information Special Report (Central Intelligence

1.2 Motivation

Given the destructive power of nuclear weapons, a state would reasonably want any sort of nuclear command and control model to properly assess reality. A false assessment leading to an unwarranted attack could wreak havoc across the globe. The situation grows increasingly complex if both countries in a two-sided conflict have automated command and control models; the addition of more models creates more room for false positives.

This analysis will define a *contradiction* as a situation in which two opposing models produce sufficiently extreme estimates assessments under the same pretense. By this definition, contradictions may arise in two scenarios: both countries believe they are weak enough to warrant a preemptive attack, or both countries believe that they are strong enough to launch a surprise first strike. For simplicity, this analysis will assume that model estimates past a predetermined threshold will result in a nuclear attack.

1.3 Research Question, Hypothesis, and Results

The possibility of a contradiction raises the central question of this analysis: Under what circumstances might two opposing nuclear command and control models produce these contradictions? Inspired by the Soviet Union's VRYAN model from the Cold War,[4] this analysis will test the hypothesis that behavioral differences between two countries can produce these contradictions.

A simulation will assess this hypothesis under various behavioral conditions, represented by the Normal, Uniform, and Cauchy distributions. The Normal distribution will represent *precise* behavior, the Uniform distribution will represent *imprecise* behavior, and the Cauchy distribution will represent *erratic* behavior.

After running 1000 iterations, the simulation confirmed the hypothesis and revealed that contradictions may arise in interactions involving erratic behavior. This result paves the way for the investigation of skewed distributions to resemble the Soviet behavior of the 1980s. This analysis also serves as a starting point for the study of scenarios in which countries adopt different launch thresholds, scoring mechanisms, and mean levels of intelligence.

Agency, February 1990), accessed October 29, 2021, https://s3.documentcloud.org/documents/2484214/read-the-u-s-assessment-that-concluded-the.pdf.

4. As future sections will describe, the model was highly inaccurate because the alarmist data fed into the model did not match the formal opinions of intelligence bodies (Board, *The Soviet "War Scare"*).

1.4 Roadmap of the Analysis

This report will assess the research question in six sections. Following the introduction, the second section will provide background on VRYAN, a Soviet command and control model from the late Cold War years. The third section will describe the limitations of the Soviet VRYAN model and propose an improved scoring scale. Then, the fourth section will introduce the simulation used to assess circumstances in which opposing models may produce contradicting assessments. The fifth section will present the results of the simulation and discuss them in the context of the research question and hypothesis. Finally, the sixth section will discuss the implications of the findings and identify areas for further exploration. Following the main analysis, the Appendix will provide supplemental information in five additional sections.[5]

2 Background: The Soviet VRYAN model in the Cold War

2.1 Motivation: The Soviet "War Scare"

The Soviet Union perceived the United States as an extreme threat in the late 1970s and early 1980s. Leadership feared that the US sought to exploit its first-strike capability and would resort to the use of nuclear weapons outside of a traditional crisis.[6] A surprise attack by the United States would greatly compress the Soviet decision-making timeframe, requiring a decision on the appropriate response in ten minutes or less.[7] These time constraints created the Soviet desire to get ahead of a potential attack.

Riding the wave of the war hysteria in Moscow, the KGB embarked on an intense intensive data collection effort, searching for signs that the US would conduct a surprise nuclear attack. After experiencing an initial influx of data in the late 1970s, the KGB told the Politburo that they needed a computer model to provide accurate assessments of the US-USSR power balance. Ultimately, this led to the development of VRYAN.[8]

5. Appendix A includes descriptions of other Soviet command and control innovations from the Cold War. Appendices B and C describe some of the shortcomings of the current model and factors to introduce more complexity. Appendix D includes additional plots to visualize the simulation results, and Appendix E describes a financial application of the command and control problem assessed in this analysis.

6. Intelligence reports indicate that the possibility of a surprise first strike by the US was improbable at the time (Board, *The Soviet "War Scare"*). However, the Soviets were not wrong to fear the US's ability to deliver a catastrophic attack on the USSR. The Joint Intelligence Committee of the UK recognized that the US could deliver such a "crippling attack" on the Soviet Union, and that fact would serve as a strong deterrent to the Soviet Government (Peter Hennessy, "The Importance of Being Nuclear: the Intelligence Picture," in *The Secret State: Whitehall and the Cold War* (London: Penguin Press, 2002), 44–77, ISBN: 0-7139-9626-9).

7. Board, *The Soviet "War Scare"*, 40.

8. Board.

2.1.1 Model overview

In the early 1980s, the KGB created a special unit to manage the VRYAN program. This computer model, whose name translates to Sudden Nuclear Missile Attack, aimed to quantify the power balance between the US and USSR. The model issued a warning if it determined that the United States achieved decisive superiority, and it operated under the premise that the US would launch an attack if it reached this level of dominance. Soviet leadership never officially tied VRYAN assessments to military operations or contingency plans, but these quick assessments would serve as guides for split-second decisions in times of crisis.[9] Because Soviet doctrine of the time emphasized the importance of preemption,[10] this analysis will assume that a warning from VRYAN would result in the launch of a preemptive attack.

The model assigned a score between 0 and 100 to the Soviet Union, fixing the United States at the maximum of 100. The score represented the percentage of strength the Soviet Union had relative to the United States. The experts tasked to develop this model believed that a score of 60 marked the threshold at which the USSR could withstand a US first strike, and anything below this threshold of 60 indicated that the Soviet Union was in a critically weak position. This analysis will refer to the predetermined VRYAN threshold as the *launch threshold*.

2.1.2 Data sources

The model consisted of a database of 40,000 weighted elements based on military, economic, and political factors defined by military and economic specialists. KGB officers constantly updated and re-evaluated the model, and Soviet leadership received formal reports with model assessments every month.[11] Data collection directives sent to Soviet field offices fell into five substantive areas: politics, economy, military, science and technology, and civil defense. A declassified CIA report stated the following high-level VRYAN collection requirements:[12]

- Plans and measures of the United States, other NATO countries, Japan, and China directed at the preparation for and unleashing of war against the “socialist” countries, as well as the preparation for and unleashing of armed conflicts in various other regions of the world.
- Plans for hostile operational deployments and mobilizations.

9. Board, *The Soviet ”War Scare”*.
10. Classified Soviet writings considered the Soviet preemption of a large-scale NATO nuclear strike as the transition from conventional to nuclear warfare in Europe (Board).
11. Board.
12. Board, 58–59.

- Plans for hostile operations in the initial stage of war; primarily operations to deliver nuclear strikes and for assessments of aftereffects.
- Plans indicating the preparation for and adoption and implementation of decisions by the NATO political and military leadership dealing with the unleashing of nuclear war and other armed conflicts.

The declassified report also stated several tasks specific to the United States:[13]

- Any information on President Reagan's "flying headquarters," including individual airfields and logistic data.
- Succession and matters of state leadership, to include attention to the Federal Emergency Management Agency.
- Information from the level of Deputy Assistant Secretary on up at the Department of State, as it was believed that these officials might talk.
- Monitoring the activities of the National Security Council and the Vice President's crisis staff.
- Monitoring the flow of money and gold on Wall Street as well as the movement of high-grade jewelry, collections of rare paintings, and similar items.

In addition to the above factors, KGB residencies abroad were instructed to collect information about possible NATO war plans, preparations for launching a nuclear missile attack against the USSR, and political decision-making leading to the initiation of war.[14]

3 Limitations and potential for improvement

While numbers carry a facade of objectivity, the strength of any mathematical model depends on its inputs and underlying construction. The rushed and paranoid development of VRYAN led to overly grim assessments. These model outputs created a self-reinforcing cycle that perpetuated the already raging war scare in the USSR.[15]

13. Board, *The Soviet "War Scare"*, 58–59.
14. Board.
15. Board.

3.1 Limitations

3.1.1 Poor construction: the shadow of WWII and a rushed timeline

The Soviet Union tasked domain experts to construct VRYAN with a focus on factors deemed important in World War II. This focus on World War II, a war in which the Soviets suffered tremendous losses, led to the overweighting of negative indicators and a skewed interpretation of reality. Despite initial KGB optimism about the power balance between the USSR and the US, the VRYAN model provided a bleak outlook for the Soviet Union.[16]

The rushed development of VRYAN further weakened the accuracy of the model's assessments. Leadership gave the KGB an incredibly short timeline to define the task, develop a plan, and begin the implementation of collection and reporting requirements. KGB officers protested, but leadership insisted on the importance of timeliness over quality. Ultimately, officers felt that the compressed timeline of the project development led to ill-conceived requirements.[17]

3.1.2 Skewed inputs: skepticism within the KGB in the face of increasing political demands

In addition to their disagreement with the project timeline, KGB officers expressed skepticism toward the VRYAN requirement in general. Many of the KGB officers concentrating on VRYAN doubted the imminence of the threat of a surprise attack by the United States. Despite doubts within the KGB about VRYAN's validity, the KGB chief made the collection of strategic military intelligence the agency's top priority.[18] Consequently, VRYAN took on a new dimension of influence.[19]

Leadership demanded increased VRYAN reporting in the face of negative VRYAN assessments. Since the model only considered negative inputs, the influx of data led to increasingly dire predictions, which then exacerbated leadership concerns.[20] Most of the top Soviet leaders of the time had formal training as engineers, so they accepted the seemingly objective VRYAN reports without hesitation.[21]

As the preoccupation with readiness heightened among political leaders, the residencies heard increasingly

16. Board, *The Soviet "War Scare"*.
17. Board.
18. This marked the first time the KGB received the order to obtain strategic military information.
19. Board, *The Soviet "War Scare"*.
20. At a special KGB conference, General Kyuchkov directly told officers that "the threat of nuclear war had reached 'dangerous proportions'" (Board, 80). Hearing this directly likely heightened the sensitivity of officers to misinterpret any little action by the West as a move to escalate to nuclear war.
21. Board.

esoteric directives. The head of KGB lauded residencies that produced high numbers of reports, which only encouraged the entry of more "innocuous information from overt sources" into the database.[22] To satisfy leadership's demands for reporting, officers submitted "willy-nilly" and "alarmist" reports that exaggerated the importance of information from unreliable sources.[23] For example, a story gleaned from a British newspaper received the label "of strategic importance" – the first time the Residency used this format in over three years – after the KGB Resident in London instructed the VRYAN officer to forward the report.[24] Not surprisingly, the selective and out-of-context nature of VRYAN reports continued to heighten fears in Moscow.

3.2 Improving VRYAN: allow for positive input

Since the model only considered negative inputs, it naturally painted a grim outlook of the power balance between the nations. A more well-rounded model would consider news of improved relations between opposing countries. It follows that enabling positive input would mark the first step to improve upon the original VRYAN model. The VRYAN model of the 1980s produced output on a scale of 0 to 100, where 100 represented the relative power of the United States and the Soviet Union could not exceed that value. However, a model that allows for positive inputs must also allow for positive outputs. The simulation in this analysis will produce output on a scale from -100 to 100, fixing the opposing country's score at 0. This scale enables each country to believe that it is up to 100 percent stronger or weaker than its opposition. Because the Soviet Union defined 60 as their original launch threshold, which fell 40 points below the assumed score of the United States, the model presented in this analysis will use -40 and 40 as launch thresholds.

When applying this updated scale to the United States and the Soviet Union, a score of -65 would signify that the Soviets have 65 percent less power than the United States. Because this score lies below the lower threshold of -40, it would indicate that the USSR should prepare for a preemptive attack. On the other hand, a score of 21 would indicate that, according to the model, the Soviet Union is 21 percent stronger than the United States. This score lies below the upper threshold of 40, indicating that the Soviet Union has an advantage over the United States but not to the extent that would warrant a surprise first strike.

22. Board, *The Soviet "War Scare"*, 82.
23. Board, 64.
24. Board, 81.

4 Simulation specifications

Thus far, the paper has described the VRYAN model of the 1980s, identified its weaknesses, and proposed an improved scale for model outputs. With that grounding, the analysis will refocus on the initial research question: In what environments might two countries with VRYAN-like models produce contradictions?

4.1 Simulation overview and layout

Drawing inspiration from the Soviet war scare of the 1980s, a simulation will test the hypothesis that behavioral differences between countries can result in contradictions. The simulation mimics behavioral differences by selecting model inputs, also referred to as *category scores*, from various distributions. The mean of a distribution represents the true relative strength of the country, and this analysis assumes that the country's intelligence reflects the true strength. While holding intelligence constant, the simulation will mimic behavioral differences by selecting category scores from various distributions. After all, the Soviet case illustrated how human behavior can distort the interpretation of intelligence.

At a high level, the simulation takes five different numbers as input, which represent the five substantive categories of the VRYAN model of the 1980s: politics, economy, military, science and technology, and civil defense. The average of the five category scores will serve as the VRYAN output. All five components of a single VRYAN score will come from the same family of distribution. However, the underlying distributions for each category may have different parameters.

It is realistic to assume that a country with erratic behavior in one respect will have comparable behavior across all categories. Using the same distribution across all categories will represent that consistency. However, a country with consistent behavior may still have relative strengths and weaknesses across the different input categories. Having different parameters for the same family of distribution will represent these relative strengths.

For example, an iteration of the model may examine an interaction in which Country A exhibits erratic behavior represented by a Cauchy distribution and Country B exhibits precise behavior represented by a Normal distribution. For this specific matchup, Country A's category scores will all come from Cauchy distributions, but the median value of the underlying distribution for the economic category may be much greater than that of the military category. Similarly, Country B's category scores will all follow a Normal distribution, but the mean parameter of the underlying distribution for the science and technology category may be much greater than the mean of the distribution for the civil defense category.

After randomly selecting a value from the specified distributions, the simulation will then compute the simple average of the five categorical scores for each distributional family. These averages will represent VRYAN outputs.[25] Each iteration of the simulation will calculate three VRYAN outputs for each country: one for the Uniform scenario, one for the Normal scenario, and one for the Cauchy scenario. After computing three output scores for both countries, the simulation will conduct pairwise comparisons of the scores. At a high level, the simulation will adopt the following structure:

1. Assign the mean parameters for each category in both countries.
2. Use the mean parameters to derive all other parameters for the various distributional scenarios.
3. Randomly select a category score from each of the specified distributions.[26]
4. Average the five category scores to produce a VRYAN output for each country and each distributional scenario.
5. If the value of the VRYAN output exceeds 100 or -100, round to the nearest 100 to comply with the VRYAN output scale.[27]
6. Compare the output of each scenario in Country A to the output of each scenario in Country B.
7. Classify any of the pairwise comparisons as a contradiction if the outputs both lie below -40 or both lie above 40.[28]
8. Repeat the above steps 1000 times in total.
9. Compute the relative frequency of contradictions for each of the six pairings of distributions.

25. Since the VRYAN output is an unweighted average, the label for each categorical score will not affect the final output. For the sake of this exercise, we can just consider the categories numbered with integers 1 through 5.

26. Since there are 5 categories, 3 distributional scenarios, and 2 countries with their separate models, each iteration of the simulation will have $5 \cdot 3 \cdot 2 = 30$ separate distributions and input scores.

27. Due to the infinite support of the Normal and Cauchy distributions, it is possible to draw category scores greater than 100 or less than -100. If this were to happen, the simulation will allow for extreme values within categories. However, the VRYAN output will round the numbers down to 100 or -100 if the averaged categorical scores exceed these limits.

28. This is the re-scaled equivalent of the Soviet-determined threshold of 60 in the VRYAN model of the 1980s. Their threshold of 60 was 40 percentage points lower than the fixed baseline score of 100, so the thresholds of this updated model will lie 40 points above and below the fixed baseline score of 0.

4.2 Introducing notation

The use of mathematical notation will help to describe the selected distributions and their parameters. Define $X_{j,k}$ as a random variable that represents Country j's input score for category k for $j \in a, b$, and $k \in 1, 2, 3, 4$. Each of these random variables will have a corresponding mean, $\mu_{j,k}$, to represent the country's true relative strength in that category. The simple average of Country j's category scores will represent the VRYAN output, $V_j = \sum_{k=1}^{5} \frac{X_{j,k}}{5}$.

For simplicity but without loss of generality, the model will assume Country A is the stronger of the two countries. For each category $k \in 1, 2, 3, 4, 5$, the simulation assigns a random value to $\mu_{a,k}$ from a $Unif(0, 100)$ distribution. Once the simulation determines the true relative strength of Country A, it will define Country B's true strength in category k as $\mu_{b,k} = -\mu_{a,k}$. For example, if $\mu_{a,k} = 40$, then Country A is k 40 percent stronger than Country B in that scoring category. Then, the simulation will assign $\mu_{b,k} = -40$, which indicates that Country B is 40 percent weaker than Country A in category k.

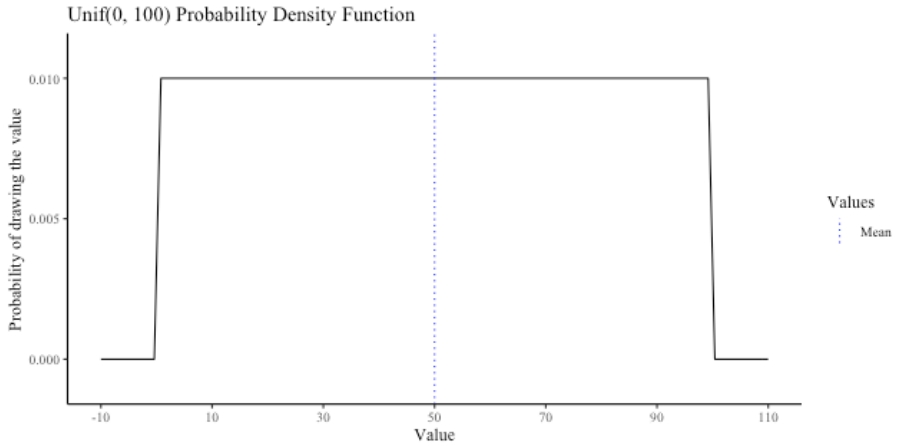

Figure 1: Probability density function to generate the $\mu_{a,k}$ values

The simulation will then draw category scores from distributions centered at the assigned values of $\mu_{j,k}$. The randomness of drawing from a distribution allows for different behaviors to glean unique interpretations of the same baseline intelligence. As an illustration that builds upon the above example where $\mu_{a,k} = 40$, Country j could still feasibly have a category score of $X_{a,k} = 72$. In this situation, Country A is only 40 percent stronger than Country B in category k, but they misinterpreted the intelligence and instead believe that they are 72 percent stronger than Country B in category k. This means that Country A viewed the intelligence in an overly optimistic light.

4.3 The Imprecise Scenario: Uniform Distribution

In this simulation, the Uniform distribution will represent imprecise but not erratic behavior. This scenario is imprecise because it can take on any value between the two endpoints with equal probability. That is, it does not assign a greater probabilistic weight to values closer to the true measure of center. However, it does not represent erratic behavior because scores from this distribution are guaranteed to lie between the upper and lower bounds. The mean, or $\mu_{j,k}$, lies at the center of the distribution. Figure 2 illustrates a scenario where $\mu_{a,k} = 50$ and the lower and upper bounds of the distribution lie at 25 and 75, respectively.

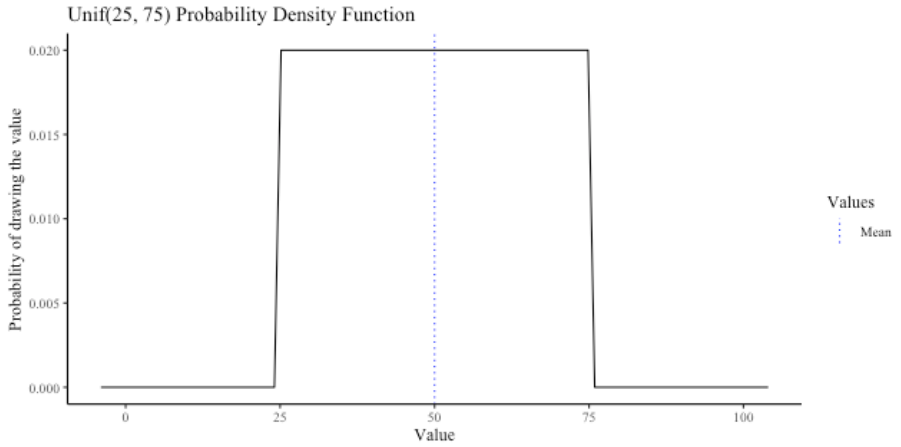

Figure 2: Probability density function for a Uniform distribution with lower bound 25 and upper bound 75

4.3.1 Defining the endpoints for the Uniform distribution

The Uniform distribution takes two parameters, the upper and lower bound, which lie equidistant from the mean. To determine these parameters, the simulation will randomly assign the upper bound and then solve for the lower bound using the predetermined mean parameter. Because the VRYAN output has an upper bound of 100, the simulation will determine the upper bound of $X_{j,k}$ by sampling a value from a $Unif(\mu_{a,k}, 100)$ distribution. This guarantees that scores generated from a Uniform distribution will not exceed the maximum value of the outputs. Then, the simulation will derive the lower bound from the distance between the assigned upper bound and $\mu_{j,k}$.

The distribution for Country B will reflect across the negative scale, so the simulation will parameterize the Uniform scenario for $X_{b,k}$ as $X_{b,k} \sim Unif(-b_a, a_a)$, where a_a and b_a represent the lower and upper bound of Country A's distribution. The two countries have perfectly symmetric distributions and parameters, but their selected category scores will likely differ in magnitude due to the random selection of a score from each

distribution.

4.4 The Precise Scenario: Normal Distribution

The Normal distribution represents precise behavior. The distribution centers itself around the mean, and values closer to the mean have a higher probability of appearing as a randomly generated score. The below figure illustrates the probability density function of a Normal distribution with a mean of 50 and a standard deviation of 20.[29] While the Normal distribution does not have a defined upper and lower bound, a randomly selected value will most likely fall within one to three standard deviations from the mean.[30]

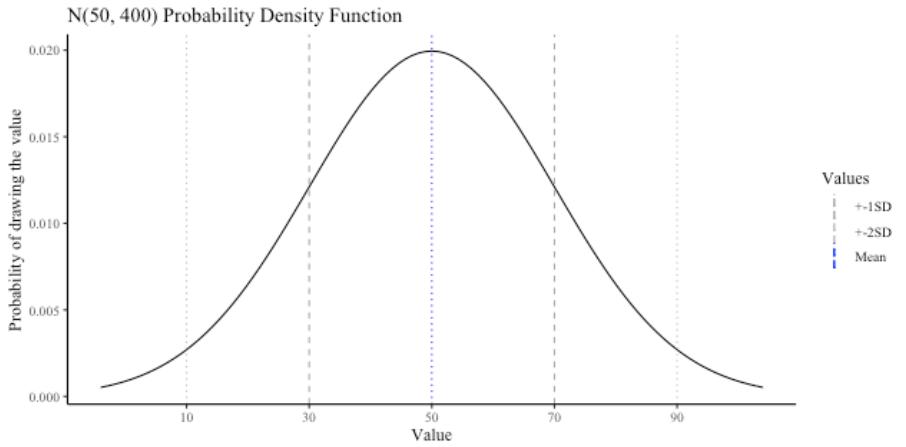

Figure 3: Probability density function for a Normal distribution with a mean of 50 and a variance of 400

4.4.1 Defining the variance for the Normal distribution

The Normal distribution has two parameters, the mean and variance, represented by μ and σ^2. Because the simulation already determined separate $\mu_{j,k}$ values for every category in a given iteration, it will only need to determine a value for $\sigma^2_{j,k}$ before selecting the category scores from Normal distributions.

The simulation utilizes the endpoints from the Uniform scenario to calculate the variance. The variance of a Uniform distribution is $\sigma^2 = \frac{(b-a)^2}{12}$, so the simulation will set the variance of the Normal distribution equivalent

29. The 400 displayed as a parameter is the variance of the distribution, which is defined as σ^2. The standard deviation is σ, which is equivalent to the square root of the variance.

30. With these specified parameters, a randomly drawn number from this distribution will lie between 30 and 70 – within one standard deviation from the mean – with a probability of 0.68. A randomly drawn value from this distribution will lie between 10 and 90 – within two standard deviations from the mean – with 0.95 probability. Finally, a randomly drawn value from this distribution will lie between -10 and 110 – within three standard deviations from the mean – with 0.997 probability.

to this value. By definition, variance is a non-negative quantity, so both Country A and Country B will have the same variance. Therefore, the variables will take on the following distributions under the Normal scenario:[31]

$$X_{a,k} \sim (\mu_{a,k}, \frac{(b_a - a_a)^2}{12})$$
$$X_{b,k} \sim (\mu_{b,k}, \frac{(b_b - a_b)^2}{12})$$

4.5 The Erratic Scenario: Cauchy Distribution

The Cauchy distribution will represent erratic behavior. The heavy tails of the Cauchy distribution allow for extremely alarmist or overly optimistic results from any given piece of intelligence. In a sense, the Cauchy scenario can mimic the extremely alarmist reports filed from relatively dull Soviet intelligence. However, the Cauchy distribution is symmetric around the median, so this high variability will occur in both directions with equal probability.

The red curve in Figure 4 displays the probability density function of a Cauchy distribution with shape and scale parameters of 50. For comparison, the gray curve displays the probability density function for a comparable Normal distribution.[32] Notice that extreme x-values have a much greater probability under the Cauchy distribution than the Normal distribution:[33]

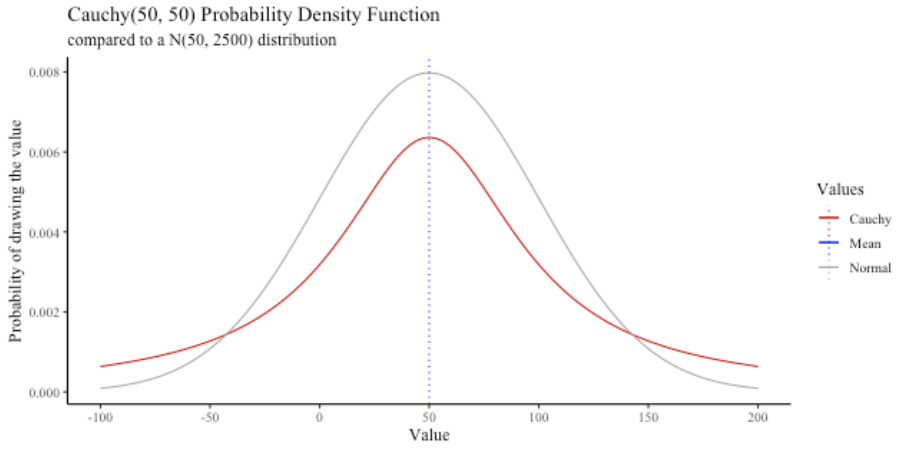

Figure 4: Probability density function of a Cauchy distribution with a shape of 50 and a scale of 50

31. a_j and b_j represent the lower and upper bound of the Uniform distribution for Country j

32. In this context, *comparable* means that these Cauchy and Normal parameters follow the same relative scale used in this simulation. Essentially, these exact two curves could appear as the erratic and precise scenarios in an iteration with a mean value of 50 and a variance of 2500 for Country A.

33. The exact spread of the graph and heaviness of the tails will depend on the scale parameter. Greater scale parameters correspond to increased variability.

4.5.1 Defining the shape and scale for the Cauchy distribution

The Cauchy distribution takes two parameters, x_0 and γ, which represent the shape and scale of the distribution. The Cauchy distribution has no mean or variance, but the shape parameter, x_0, is equivalent to the median of the distribution. To remain consistent with the other distributional scenarios, the simulation will define the shape parameter as the previously determined mean value. That is, $x_{0j} = \mu_{j,k}$ for Country j.[34] The scale parameter represents the spread of the data, so the simulation will use the standard deviation – which equals the square root of the variance – from the Normal scenario.[35]

5 Simulation results

The simulation confirmed the hypothesis that different behavioral conditions can result in contradictions as defined in this analysis. However, contradictions only occurred under specific conditions. Under the controlled circumstances of the simulation, contradictions only appeared if at least one of the countries exhibited the erratic behavior of the Cauchy distribution. Since this simulation held all other variables constant, this indicates that erratic behavior has the power to cause these contradictions. Figure 5 displays the proportion of contradictions across the 1000 iterations of each behavioral matchup.

Direct comparison of the above proportions allows one to gauge the effect of different behaviors on the frequency of contradictions. Pairings of two countries with erratic behavior resulted in contradictions most frequently, with contradictions occurring in 6% of iterations. Then, contradictions appeared in matchups between an erratic country and an imprecise country in 3.3% of iterations. Finally, pairings of an erratic country with a precise country produced contradictions in 2.8% of the iterations. The remaining pairings did not produce any contradictions across the 1000 iterations of the simulation.

34. The Cauchy distribution has no mean or variance, but the shape parameter, x_0, represents the center of the distribution. For the sake of consistency with the other distributional scenarios, it will assign the value of $\mu_{j,k}$ parameter to $x_{0j,k}$.

35. While there is no way to analytically solve for the scale parameter, one can approximate the scale as $\hat{\gamma} = median(|X|)$, which is the median of the absolute values of the data. Since the simulation is producing data from the specified distribution rather than parameterizing the distribution from already-existing data, it is impossible to determine the scale parameter by taking the median of data that does not exist. Instead, the simulation must define the scale parameter before producing data.

The simulation will use standard deviation for several reasons. First, both the standard deviation and the scale parameter must be non-negative. Second, using the variance as the scale parameter would cause the Cauchy distribution to have incredibly erratic behavior for cases that have a large variance. This erratic behavior leads to a high proportion of outputs far beyond the range of the -100 to 100. Using the square root of this value, however, creates much more manageable categorical scores.

Scenario	Contradictions
Both Cauchy	0.06
Uniform Cauchy	0.033
Normal Cauchy	0.028
Both Uniform	0
Both Normal	0
Uniform Normal	0

Figure 5: The proportion of contradictions across 1000 iterations of each behavioral matchup.

6 Conclusion

Holding all other variables constant allowed for the study of the causal effect of behavior on the frequency of contradictions. While this simulation did not yield any shocking results, it reveals that differences in behavior alone can produce these contradictions. Two countries with symmetric levels of intelligence, matching launch thresholds, and identical scoring systems can still experience contradictions if at least one of the two countries exhibits erratic behavior.

With that said, several improvements would strengthen the statistical validity and rigor of the simulation. For example, none of the considered distributions exhibit behavior that skews in a single direction, but this would more closely resemble the alarmist behavior of the VRYAN officers in the 1980s. Additionally, future versions of this simulation may reconsider the assignment of parameters.[36]

More broadly, the research question asks about what *circumstances* might two opposing models produce these contradictions. Due to the complex nature of any command and control model, a simulation could easily force the desired results by manipulating several variables at once. To engage in a more meaningful analysis, the hypothesis of this paper focused only on the underlying distributions of the inputs. Future work on the research question should expand upon the distributional focus of this analysis. Lifting many of the safeguards held constant in this simulation would allow for the study of model interactions where the countries have different levels of intelligence, variance of inputs, and scoring mechanisms.[37] As a further extension, simulations could

36. See Appendix B for a list of possible improvements to the current simulation.

37. See Appendix C for a list of approaches to add complexity and nuance to the present analysis.

explore the integration of a VRYAN-like model with launch systems.[38] Despite leaving these areas unexplored, the work conducted in this simulation marks an important first step in exploring two-way interactions between command and control models.

38. Appendix A.2 and Appendix C.5 propose methods to accomplish such a task

A Additional Soviet command and control automation efforts

This analysis focused on the Soviet Union's VRYAN model, but the USSR created numerous other systems to streamline their nuclear command and control processes.

A.1 Monolit: transmitting code words via radio

Early Cold War communications relied on the *Monolit* system, which transmitted orders to missile commanders through cable and radio with a system of code words. Prepared in advance for the event of nuclear war, paper packets contained instructions to change frequencies and call signals for all Soviet radio systems. Upon receiving a code word through the Monolit system, crews had to open the packet and introduce the new radio information everywhere.[39]

This primitive system had numerous shortcomings. In drills, nervous officers struggled to open the packets with scissors. Additionally, the Monolit system did not allow officers to cancel or recall orders. This created the need for a more streamlined communication system.

A.2 Signal, Signal-M, and Cheget: command-to-launch systems

Created in 1967, the *Signal* system marked the first step toward automating Soviet command and control processes. The system could transmit thirteen fixed commands to troops from headquarters, and it allowed for the cancellation of orders. While Signal marked an improvement from the paper packets, it still relied on the troops to operate the weapons upon receipt of orders rather than directly commanding the weapons.

In the mid-1970s, *Signal-M* marked the second stage of automation. This new system reached all levels of decision-making and marked the first introduction of a push-to-launch button in Soviet nuclear command and control. Leaders in the Kremlin still did not have sole control over the launch apparatus; rather, the military branches kept the launch apparatus and shared control between generals and political leaders.[40]

In 1985, the *Cheget* system marked an upgrade to a more modern system resembling a nuclear football. This system still lacked the power to launch weapons itself. Rather, it plugged into a communications network that allowed the general secretary to grant permission to the military to issue a launch. Upon receipt, these

39. David E. Hoffman, "The Dead Hand," in *The Dead Hand: the untold story of the Cold War arms race and its dangerous legacy* (New York: Doubleday, 2009), 146.

40. Hoffman, 148.

permissions turned to direct commands. Following authentication, these direct commands transformed to launch commands and were sent to the missiles.[41]

A.3 Perimeter: streamlining the chain of command

The *Perimeter* system marked the last official innovation in the launch process during the Cold War years. The system automated the flow of launch orders through the chain of command, with the soldiers in a bunker serving as the final link. Soviet planners and theorists could not reliably predict when leaders should launch a nuclear missile attack to optimize the survival of the nation, and the declining health of Chernenko meant that his physical and cognitive capacity to launch might fail. As a response to these seemingly impossible circumstances, Soviet designers created a command system that allowed Chernenko to opt out of decision-making and pass the burden to someone else. Essentially, this took the weight of the decision off the ailing leader and placed it on a more competent survivor within the chain of command.[42]

Perimeter assessed three conditions, that if true, would prompt junior officers in concrete bunkers to launch a nuclear attack: the system was activated, leaders were dead, and bombs were detonating. After successfully executing a test, Perimeter was put on combat duty in 1985.[43] The Perimeter system still relied on a human component: the officers in the bunker. These officers could either buckle under pressure and decide not to wreak havoc on the globe, or they would stick to their training and launch without a second thought. The second option, where the officers do not think about their actions, led to discussions about a more dystopian system.

A.4 Dead Hand: a fully automated launch system

Perimeter birthed the idea of the *Dead Hand* system, an entirely automated version of Perimeter. Planners aimed to create an automated retaliatory system that could order an attack without any sort of human control, essentially enabling a nuclear launch even when all other hands were dead. Soviet designers and leaders considered this system in the 1980s, but they eventually rejected the idea as "madness" and abandoned the project.[44]

41. Hoffman, "The Dead Hand," 149.
42. Hoffman.
43. Hoffman, 154.
44. Hoffman, 152.

B Limitations of the current simulation

As discussed throughout the analysis, the simulation presented in this analysis holds many variables constant, only varying the underlying distribution of category scores for each country. This hardly explores the breadth of ways opposing command and control models could produce contradictory outputs. However, this model will require further refining before exploring the impact of changing means, thresholds, and model constructions. The following section will outline potential improvements for a more refined version of the current simulation.

B.1 Selecting $\mu_{j,k}$ from a Normal distribution rather than Uniform

Each iteration began by determining each category's mean parameter from a $Unif(0, 100)$ distribution. With this underlying distribution, every variable will have an expected power balance of +50 percentage points for Country A relative to Country B. That is, $E(\mu_{a,k}) = 50$ for $k \in 1, 2, 3, 4, 5$. Because the simulation used a default launch threshold of 40, this meant that, on average, the simulation would result in an attack on Country B from Country A.

Realistically, the USSR did not have an equal probability of taking on any score from 0 to 100 on its original scale; their VRYAN output never dropped below the threshold of 60. While this simulation did not aim to replicate the USSR model, but it should represent a realistic interaction between two countries. Future iterations should assign a greater probability to $\mu_{j,k}$ values closer to 0 since two countries are more likely to be relatively balanced than they are to be on the brink of nuclear war. Future iterations could draw the $\mu_{a,k}$ parameter from a Normal distribution centered at 0.[45] The simulation would still assign $\mu_{b,k} = -\mu_{a,k}$ to reflect symmetry in the true strength of each country.[46]

B.2 Assigning parameters for the Normal and Cauchy scenarios

Methods to assign parameters other than $\mu_{j,k}$ would also benefit from more statistical sophistication. The current method of assigning the variance and scale parameters depends on the value of $\mu_{j,k}$, but a better model

45. This poses new issues with the support. A Uniform distribution established an upper bound at 100. For this Normal distribution, the simulation could set the variance as $\sigma^2 = (\frac{100}{3})^2$, which would ensure that 99.7% of the values fell between 100 and -100. If the selected $\mu_{j,k}$ exceeded -100 or 100 in magnitude, the simulation could simply round it off to those bounds.

46. Under this proposed setup, Country A is no longer guaranteed to have the higher mean of the two countries, which comes with a new set of weaknesses. Realistically, category scores should be correlated within each country. The current setup means that a country that is stronger in one respect is stronger in all respects, but the proposed setup would make it possible for a country with a very unstable economy to have a relatively strong political balance.

would have independent measures of center and spread.[47] Ideally, more time and resources would go toward determining the most appropriate way to assign the Normal variance and the Cauchy shape and scale. Further statistical research would allow for a systematic selection of distributions and parameters.

B.3 Including at least one skewed distribution

Ideally, a simulation like the one presented in this analysis would include a skewed distribution in addition to the heavy-tailed Cauchy. For example, the distribution of the USSR's scores would have taken on a right skew since they exhibited greater sensitivity to negative information and interpretations. Essentially, they would assign lower scores with greater probability than they would assign scores closer to their $\mu_{j,k}$ value. The fat tails of the Cauchy distribution represent volatility in both directions, but it fails to capture the behavior of a country that is consistently over-optimistic or pessimistic.

47. In the current setup, $\mu_{j,k}$ will be correlated with the $\sigma^2_{j,k}$ and $\gamma_{j,k}$ parameters. the endpoints of the Uniform distribution determine the variance parameter of the Normal distribution. The scale parameter of the Cauchy distribution is equal to the standard deviation – or the square root of the variance – of the Normal distribution. Currently, the upper bound of the Uniform distribution must lie between $\mu_{a,k}$ and 100. As a result of this, higher values of $\mu_{a,k}$ will be associated with shorter distances between the mean and the upper bound. A shorter distance between endpoints for higher values of $\mu_{a,k}$ means that these higher values of $\mu_{j,k}$ will be associated with smaller variances. A smaller variance will then lead to a smaller standard deviation and therefore a smaller scale parameter.

C Factors to consider for a more complex simulation

C.1 Systematically vary parameters within iterations

This simulation only varied the distribution and held all other variables as constant or symmetric between the two countries. Future versions of this simulation should explore the impact of changing the center and spread of each distribution. This would allow for the simulation to record the extent that conditions can differ in each scenario without producing contradictions.

C.2 Include several sophisticated scoring systems

Just as this simulation held the mean and variance constant between the two countries, it also used the same scoring mechanism for both sides. Realistically, two countries would likely develop unique methods to quantify the power balance between two countries. Moreover, it is highly doubtful that any country would allow an arithmetic mean to determine if they should launch a nuclear attack. Both countries would likely have more complex scoring mechanisms that differ from one another.

A future iteration of this project would develop more sophisticated scoring systems. Ideally, these would draw from real-life data, such as the Correlates of War dataset.[48] The project would fine-tune hyper-parameters using cross-validation and perform model selection as if nuclear war depended on it. Just as future more complex simulations would systematically vary the measures of center and spread between countries, they would also allow for countries to adopt different sophisticated scoring mechanisms.

C.3 Vary the threshold between countries

This simulation held attack thresholds constant at 40 and -40 for both countries in every iteration. Similar to how two countries would likely develop different scoring systems, two countries would probably determine unique thresholds to launch an attack. Future versions of this simulation could vary the launch thresholds between countries and iterations and examine how this impacts the proportion of contradictions.

48. Meredith Reid Sarkees and Frank Wayman, *Resort to War: 1816-2007* [in en] (Washington DC: CQ Press, 2010), accessed November 28, 2021, https://correlatesofwar.org/data-sets/COW-war.

C.4 Consider multiple classification thresholds

This simulation considered a model that produced binary results that either resulted in an attack or did not result in an attack. However, a country could consider other actions based on different levels of output. For example, a country could set thresholds for declaring war, sending troops abroad, and other steps along the path to nuclear escalation.[49] This would open the door to further study of the impact of these models on the pace of escalation more broadly.

C.5 Integrate VRYAN with Perimeter

As discussed, model assessments from VRYAN were never tied to military operational plans or contingency plans, but this analysis assumed that crossing the threshold would result in a launch. In keeping with the desire to maximize efficiency under extreme time and pressure constraints, it would make sense for VRYAN to trigger some sort of launch mechanism. Western moral imperatives would not allow for a computer to simply launch a nuclear weapon without human intervention. However, the Soviet Union developed a compromise of sorts and put it into combat in 1985. In the event of a nuclear attack, the Perimeter system allowed human actors to launch a counterattack, cancel the order, or defer to the next person in the chain of command. This automated system involved human actors at every step in the launch system.

Given the human element of Perimeter, it would not be unrealistic for VRYAN to trigger the Perimeter system if the output score crosses the specified threshold. For simplicity, this analysis focused on the output of a VRYAN-like model with the underlying assumption that crossing a particular threshold would lead to a preemptive attack or a surprise first strike. A more sophisticated simulation, however, would activate a Perimeter-like model if the VRYAN output fell beyond the specified threshold. At each node in the chain, the simulation would assign each leader a certain probability to either defer to the next person, cancel the order, or launch an attack. The order would flow through the chain until someone chose to launch an attack or cancel the order. Simulations could introduce random noise in the probabilities of making each decision to represent varying levels of emotional stress in leaders.

49. Herman Kahn, "On Escalation: Metaphors and Scenarios" [in en], in *On escalation: metaphors and scenarios* (New York: Praeger, 1965), 3–61, accessed December 2, 2021, https://www.routledge.com/On-Escalation-Metaphors-and-Scenarios/Kahn/p/book/9781412811620.

D Supplementary graphs

D.1 Distribution of outputs

The below plots display the distribution of the VRYAN outputs of each country under the different behavioral scenarios:

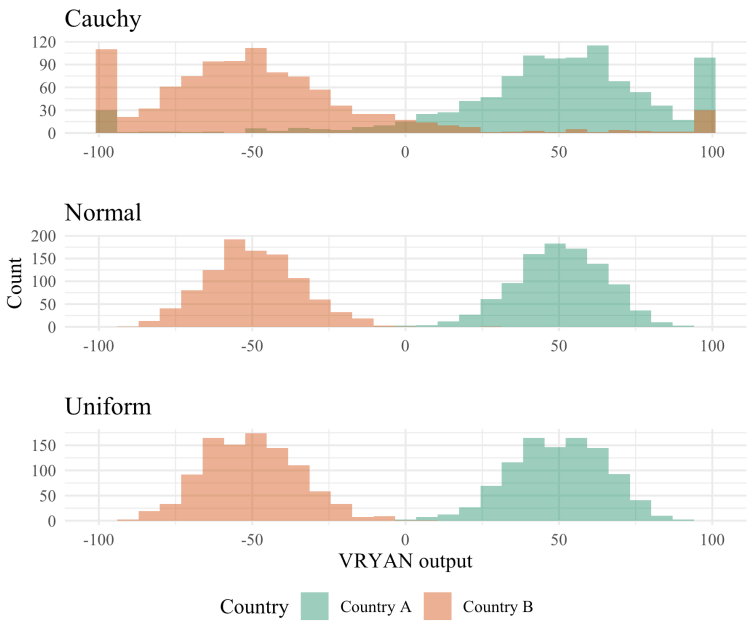

Figure 6: Distribution of VRYAN outputs under each scenario

Notice that all three of these distributions follow a somewhat bell-shaped curve. This follows from the Central Limit Theorem and the use of the simple average to determine the VRYAN output.[50] Bell shape aside, the outputs from the Cauchy scenario have much more variability than the other two scenarios, as demonstrated by the wider spread in the values. The high bars at -100 and 100 are a result of rounding any VRYAN output that exceeded the limits of 100 or -100. If the outputs were not limited to this range, then the curves would extend beyond -100 and 100 with relative symmetry in both directions.

The distributions of outputs produced from Normal and Uniform categorical scores look relatively similar, with the y-axis of the Uniform curve showing fewer values concentrated at the center. This results from the lack of precision when drawing from a Uniform distribution as opposed to a Normal distribution, and it illustrates why the Normal distribution represents precise behavior while the Uniform distribution represents imprecise behavior.

50. Richard Routledge, *Central Limit Theorem*, accessed December 5, 2021, https://www.britannica.com/science/central-limit-theorem.

D.2 Visualization of contradictions

As previously described, contradictions may arise in two scenarios: both countries believe they are sufficiently weak enough to warrant a preemptive attack, or both countries believe that they are strong enough to warrant a surprise first strike. Figure 6 illustrates this concept; a contradiction means that both countries' outputs fell on the same exterior side of the two dotted lines at the thresholds of 40 and -40. For conciseness, the below figure displays a random sample of 30 VRYAN outputs of the 1000 total outputs recorded by simulation. Also for the sake of brevity, Figure 6 only displays three of the six behavioral matchups:

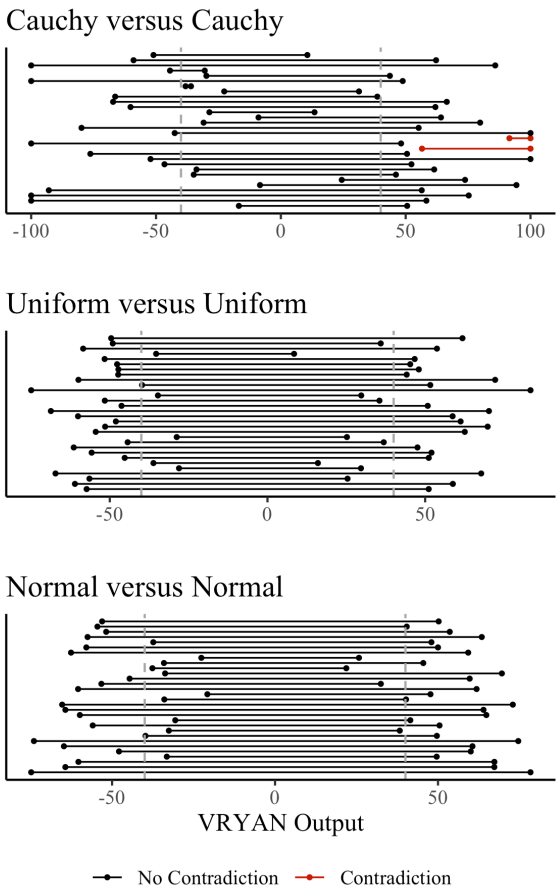

Figure 7: Random sample of 30 matchups

E Flash Crash: Similar Application in Finance

E.1 What is a flash crash?

Financial markets provide a non-military application of the concepts illustrated in this paper. The interaction of automated trading systems can cause *flash crashes* in financial markets. This occurs because high-frequency traders (HFTs) demand immediacy ahead of other market participants. The increasing use of HFTs has led to a continuous market presence, which creates liquidity imbalances. When paired with large sell orders, liquidity imbalances can lead to a liquidity-based crash, characterized by high trading volume and price volatility.[51]

Under calm conditions, HFTs do not pose issues. However, automated trades will contribute to volatility if there is already an imbalance of order flows and a directional movement of prices. In this case, HFTs act faster than the speed at which the best bid and offer queues are taken and fulfilled. The speed of these HFTs will result in a spike in trading volume, which creates an environment for a flash crash.

E.2 Flash Crash of 2010

The most notable example of a flash crash occurred on May 6, 2010. During key moments of low market liquidity, HFTs removed the last few contracts at best bids and demanded additional depth beyond what was available. The S&P 500, Nasdaq 100, and Russell 2000 collapsed and rebounded rapidly in 36 minutes, marking the largest intraday point decline in DJIA history. More broadly, stock index futures, options, ETFs, and individual stocks saw extreme volatility and spikes in trading volume.[52]

E.3 Simulating flash crashes

Due to the small number of observed flash crashes, researchers can use simulation as a tool to study and manipulate the conditions surrounding flash crashes. For example, researchers have used simulation to analyze liquidity and price variability under proposed interventions and regulatory frameworks.[53]

51. Andrei A. Kirilenko et al., "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market" [in en], *SSRN Electronic Journal*, 2011, ISSN: 1556-5068, accessed December 5, 2021, https://doi.org/10.2139/ssrn.1686004, http://www.ssrn.com/abstract=1686004.

52. Kirilenko et al.

53. Paul Brewer, Jaksa Cvitanic, and Charles R. Plott, "Market Microstructure Design and Flash Crashes: A Simulation Approach," Publisher: Routledge _eprint: https://doi.org/10.1016/S1514-0326(13)60010-0, *Journal of Applied Economics* 16, no. 2 (November 2013): 223–250, ISSN: 1514-0326, accessed December 5, 2021, https://doi.org/10.1016/S1514-0326(13)60010-0, https://doi.org/10.1016/S1514-0326(13)60010-0.

E.4 Relating flash crashes to this analysis

Two opposing HFTs can lead to volatility in trading volume, just as two opposing command and control models may lead to rapid escalation. Similarly, simulations of flash crashes allow researchers to examine the effect of different regulatory frameworks, just as this analysis studied the effects of different behaviors on the VRYAN-like model outputs.

However, flash crashes differ from nuclear escalation in one key feature. The study of flash crashes focuses on regulations and interventions to prevent such events. If two states reach the point of nuclear conflict, regulations will make little impact. When applied to nuclear escalation, a simulation will not find a clear-cut solution, but it will help to build an understanding of the implications of automation in defense and national security.

References

Board, President's Foreign Intelligence Advisory. *The Soviet "War Scare"*. Intelligence Information Special Report. Central Intelligence Agency, February 1990. Accessed October 29, 2021. https://s3.documentcloud.org/documents/2484214/read-the-u-s-assessment-that-concluded-the.pdf.

Brewer, Paul, Jaksa Cvitanic, and Charles R. Plott. "Market Microstructure Design and Flash Crashes: A Simulation Approach." Publisher: Routledge _eprint: https://doi.org/10.1016/S1514-0326(13)60010-0, *Journal of Applied Economics* 16, no. 2 (November 2013): 223–250. ISSN: 1514-0326, accessed December 5, 2021. https://doi.org/10.1016/S1514-0326(13)60010-0. https://doi.org/10.1016/S1514-0326(13)60010-0.

Grote, Thomas, and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare" [in eng]. *Journal of Medical Ethics* 46, no. 3 (March 2020): 205–211. ISSN: 1473-4257. https://doi.org/10.1136/medethics-2019-105586.

Hennessy, Peter. "The Importance of Being Nuclear: the Intelligence Picture." In *The Secret State: Whitehall and the Cold War*, 44–77. London: Penguin Press, 2002. ISBN: 0-7139-9626-9.

Hoffman, David E. "The Dead Hand." In *The Dead Hand: the untold story of the Cold War arms race and its dangerous legacy*, \autocite{presidentsforeignintelligenceadvisoryboardSovietWarScare1990}. New York: Doubleday, 2009.

Imai, Kosuke, and Zhichao Jiang. "Principal Fairness for Human and Algorithmic Decision-Making," January 2021. Accessed November 28, 2021. https://imai.fas.harvard.edu/research/fairness.html.

Kahn, Herman. "On Escalation: Metaphors and Scenarios" [in en]. In *On escalation: metaphors and scenarios*, 3–61. New York: Praeger, 1965. Accessed December 2, 2021. https://www.routledge.com/On-Escalation-Metaphors-and-Scenarios/Kahn/p/book/9781412811620.

Kirilenko, Andrei A., Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market" [in en]. *SSRN Electronic Journal*, 2011. ISSN: 1556-5068, accessed December 5, 2021. https://doi.org/10.2139/ssrn.1686004. http://www.ssrn.com/abstract=1686004.

Routledge, Richard. *Central Limit Theorem*. Accessed December 5, 2021. https://www.britannica.com/science/central-limit-theorem.

Sarkees, Meredith Reid, and Frank Wayman. *Resort to War: 1816-2007* [in en]. Washington DC: CQ Press, 2010. Accessed November 28, 2021. https://correlatesofwar.org/data-sets/COW-war.

Sayler, Kelley M, and Daniel S Hoadley. "Artificial Intelligence and National Security" [in en], November 2020, 43.