Probabilistic Counterfactual Analysis

September 25, 2025

> "[The question of] what might have happened if, for example, Bismarck had not decided to make war [in 1886] is by no means an 'idle' one... It does indeed bear on something decisive for the historical moulding of reality, namely, on what causal *significance* is properly attributed to this individual decision in the context of the totality of infinitely numerous 'factors.' ..." (Weber 1949, p. 164, emphasis in original)

In this paper I offer a systematic and disciplined way to probabilistically evaluate most likely outcomes in a counterfactual world by weighing the relevant historical evidence from the real world. In introducing this novel method, probabilistic counterfactual analysis, I offer a generic five-step process to analyze a counterfactual world. The five steps are: (1) observe the real world to confirm that X appears before Y, (2) create a counterfactual world in which X is absent through a *boundedly arbitrary rewrite*, (3) construct a decision tree by identifying decision points through a close reading of real world history, (4) convert the decision tree into a probability tree to calculate how confident we would have to be at each decision point to conclude whether Y or no Y was more likely under the counterfactual premise, and (5) weigh the historical evidence to make an assessment about whether Y or no Y was more likely. In addition to being systematic and disciplined, the novelty of my method is its *explicitly probabilistic* approach to analyzing counterfactual worlds. Existing approaches to counterfactual worlds are interested in whether a given *counter-to-the-facts* premise (often called the antecedent) would have led to a given *counter-to-the-facts* outcome (often called the consequent). Because the way they approach this question is yes-or-no

regarding a particular consequent, I consider these existing approaches to be a *deterministic* approach to counterfactual analysis. In contrast, my *probabilistic* approach considers two consequents - which can also be called potential outcomes - Y and no Y and evaluates the *probability* of one potential outcome relative to the *probability* of the other potential outcome in the counterfactual world.

One particularly fruitful area that this method could be applied to is studying the role of leaders in international relations. When it comes to establishing whether leaders were dispensable or indispensable, Greenstein (1992, p. 123-124) explicitly calls for counterfactual analysis: "Establishing whether [an] individual's personality shaped the outcome calls for a ... complex reconstruction that asks whether the situation of the actor in question would have imposed the same course of action on anyone who might plausibly have occupied that individual's position. This calls for examining not only the psychology of the individual in question, but also the historical context, including the other significant actors and their claims, demands, perceptions, and personal qualities" - in other words, an exercise in counterfactual reasoning. "This is the only available alternative in analyses of single events to the quantitative analysis that would be called for if data existed on large numbers of comparable episodes. Counter-factual reasoning is not falsifiable, but it can be systematic. To be so it must be explicit and addressed to bounded questions" (Ibid., p. 124). Jervis (2013, p. 158-160) also explicitly recommends counterfactual analysis in evaluating whether leaders "mattered" or not. As no comparison will be perfect and there will only be few appropriate ones to make, counterfactuals can also be used, "although the rules of this game are easy to dispute. ... the claim that the characteristics of an individual leader matter implies that the state's policy would have been significantly different had someone of different characteristics been in power. This is always true if we make an extreme mental substitute - had Adolf Hitler been president, American foreign policy would have been enormously different. But this could not occur unless the United States had been transformed into a fundamentally different country. More illuminating counterfactuals would replace a president with one of

his or her main challengers or with someone who was a plausible candidate but chose not to run" (Ibid., p. 160). Probabilistic counterfactual analysis as a method responds to these suggestions, and takes their advice and warnings seriously; probabilistic counterfactual analysis is systematic, addressed to bounded questions, and focuses on more illuminating, i.e. "plausible," counterfactuals.

The rest of this paper is organized into five parts. First, I offer a review of existing approaches to counterfactual analysis in international relations and discuss the cotenability requirement. Second and most importantly, I introduce probabilistic counterfactual analysis as a method. Third, I discuss the novelty of my approach compared to existing approaches to counterfactual analysis. Fourth, I demonstrate the utility of the approach by applying it to the question of Neville Chamberlain's role in WWII (Khong 1996). Fifth and finally, I conclude by discussing what kinds of political science questions the method may be applied to.

Counterfactual Analysis: Existing Approaches

A counterfactual is a "subjective conditional in which the antecedent is known or supposed for purposes of argument to be false" (Tetlock and Belkin 1996, p. 4). Counterfactuals are an explicit and implicit part of causality as it is studied in political science, as any causal statement (If X, then Y) explicitly or implicitly generates a counterfactual statement (If no X, then no Y) that must also be true if the original causal statement is to be true (Brady 2008; Lewis 1973).[1] Before introducing probabilistic counterfactual analysis, I offer a review of counterfactual analysis as a method and how counterfactual analysis has been used in the study of international relations thus far.

Fearon (1991, p. 194), among many others, points out the prevalence of counterfactuals in political science: "Counterfactuals and the counterfactual strategy of hypothesis testing play an important but often unacknowledged and underdeveloped role in the efforts of political

[1]For more on the causal logic of necessary and sufficient conditions and how necessary condition counterfactuals can be used to explain the causes of individual events, see Levy and Goertz (2007).

scientists to assess causal hypotheses. ... any *non*experimental research that makes causal claims, ... must confront counterfactuals in the form of key assumptions or in the use of hypothetical comparison cases. Particularly in small-N research ... counterfactual thought experiments [are] a necessary means for serious justification of causal claims." Counterfactual analysis differs from cross-case comparisons in the kinds of risk they run and how they assess support for causal hypotheses. The main risk in counterfactual analysis is "how can we know what would have happened with any degree of confidence?" (p. 174).[2] In counterfactual analysis, instead of causal hypotheses being supported by frequencies of associations (i.e., correlations), "support for a causal hypothesis in the counterfactual strategy comes from *arguments* about what would have happened. These arguments are made credible (1) by invoking general principles, theories, laws, or regularities distinct from the hypothesis being tested; and (2) by drawing on knowledge of historical facts relevant to a counterfactual scenario" (p. 176, emphasis in original).[3] Fearon's key takeaway is not that counterfactuals should be used more or should not be used at all, but that "researchers who use counterfactual argument to support causal hypotheses should be methodologically aware of what they are doing and should make their counterfactual arguments as *explicit* and *defensible* as they can" (p. 170, emphasis added). The purpose of probabilistic counterfactual analysis as a method is to do just this: to make single-case studies more methodologically self-aware by forcing analysts to make explicit claims about what would have been different at each decision point

[2]In considering large-N studies that are not experimental, Fearon argues that a counterfactual proposition is smuggled into the assumption that the value of the independent variables and error terms are uncorrelated. In other words, for regression analysis to be credible, the counterfactual proposition - "If the cases in the sample had assumed different values on the independent variables, the contents of the error term would not have differed systematically" - must also be true (p. 175).

[3]A second key difference between counterfactual and actual case strategies is that in actual case strategies, the *relative causal weight* of different independent variables can be estimated by extracting relevant frequencies and magnitudes from the sample of real world cases. In contrast, in the counterfactual analysis, because there is no concrete frequencies or magnitudes, every time a new variable is introduced that could affect the potential outcome, we must add more counterfactual cases. And the third key difference between the two strategies is the *precision of estimates*. In large-N studies, it is possible to estimate how confident we are that this given independent variable has this much of a causal effect. In counterfactual analysis, there is no such procedure to estimate how confident we can be when we say a given independent variable had this much of a causal effect. "All depends instead on the plausibility of arguments about what would have happened. ... *arguments about the relative importance of possible causes become arguments about the relative plausibility of different counterfactual scenarios*" (p. 178, emphasis in original).

in the counterfactual world, and to bring to bear relevant evidence to defend their claims about what would have most likely happened at each decision point in this counterfactual world.

In another piece on counterfactuals, Fearon (1996) stresses the impossibility of knowing "what would have happened" using an analogy of historical processes to cellular automata. Fearon's point is that even with very basic rules about how cellular automata change from period $t = n$ to $t = n + 1$, it is impossible to predict what exactly would happen several iterations later, $t = n+x$. He writes, "It is in the nature of automatonlike processes that one may not be able to predict what will happen several steps ahead without explicitly 'running' the system. Attempting to proceed by deduction from rules plus initial conditions will take one only a very short distance forward before complexity overwhelms the effort. Moreover, this can be true *no matter how good our knowledge of the transition rules or theories used to draw out the consequences of a particular counterfactual antecedent.*" (p. 49, emphasis in original). Here it is important to underscore that counterfactual analyses have very different "inputs" depending on the question. If the relevant input into the counterfactual analysis is theoretical arguments, what are analogous to the basic rules in cellular automata, he is right about the near-impossibility of "what would have happened." However, depending on the question, if the relevant input into the counterfactual analysis is historical evidence and not theoretical arguments, I posit that it is possible to get a better idea of "what would have happened," and even put some bounds around how confident we are that this is "what would have happened."

Though I would posit that for narrowly defined questions in data-rich cases it is possible to get a better understanding of "what would have happened" in a counterfactual world, I agree with Fearon about the importance of proximity between the counterfactual antecedent and hypothesized consequents: "detailed counterfactual scenarios will have a chance at being rendered plausible only if the proposed causes are temporally and, in some sense, spatially quite close to the consequents" (p. 50). In discussing how historical processes are stochastic,

Fearon foreshadows my approach when he argues, "The point is that if we think historical processes evolve stochastically - or stochastically for all we can possibly know - then counterfactual antecedents do not imply determinate paths for counterfactual scenarios, as we often assume in sketching them. Rather, *changing a factor counterfactually implies a probability distribution over many counterfactual paths, in which the evidence provided by the actual world that did occur may not be relevant for saying what would have occurred in unprecedented counterfactual situations*" (p. 51, emphasis added). While Fearon foreshadows the idea of thinking about counterfactual analysis in terms of paths and probabilities, I am the first to formalize this approach. And finally, Fearon in this article introduces the distinction between conceivable causes - "they could conceivably have happened according to the 'rules of the game' as we understand them" - and miracle causes - "we imagine their counterfactual occurrence as resulting from an intervention from outside the system (the hand of God, as it were)" (p. 60). Fearon suggests that "we should only consider counterfactual antecedents that might affect the consequent while leaving the rest of the world otherwise similar" (p. 67), though he admits the usefulness of miracle causes when the analyst is "treating the specific event to be explained as an instance of a class of events" (p. 61).

Tetlock and Belkin (1996, p. 6) begin their edited volume on counterfactuals by noting its prevalence in causal arguments: "Try though we do to control statistically for confounding variables in large-N multivariate studies or to find matching cases in comparative designs or to search for the signature of hypothesized causes in process-tracing studies, the potential causes are simply too numerous and too interrelated in world politics to permit complete escape from counterfactual inference." They go on to identify five styles of counterfactual argumentation: idiographic, nomothetic theory-testing, idiographic-nomothetic synthesis, pure thought experiments (logical proofs and computer simulations),[4] and mental simulations of counterfactual worlds.[5] The first three styles are the most directly relevant to positivist

[4]Pure thought experiments, through logical proofs and computer simulations, can be used to "reveal previously hidden contradictions or ambiguities in the logical structure of the causal arguments that others have advanced" (p. 12).

[5]Mental simulations of possible worlds similarly "highlight critical contradictions and ambiguities in one's

political science arguments. Idiographic counterfactuals are used to "explore 'possibility-hood' - whether history had to unfold as it did" (p. 7). Idiographic counterfactuals focus on how different the world could have been in response to small changes, are constrained by the "minimal-rewrite-of-history" rule, are a useful corrective to deterministic theory, and are necessary to pass moral judgment on leaders (p. 7-8). Nomothetic theory-testing counterfactuals "usually show little or no concern for the plausibility of switching the hypothesized counterfactual antecedent on or off in any given context" (p. 8). These employ what Fearon calls miracle rewrites, and the purpose is not to understand history better but to understand the "logical implications of a theoretical framework better" (p. 9). Another way to distinguish between the two is that idiographic counterfactual causal argument concerns "the impact of a particular belief, person, or policy" while a nomothetic counterfactual causal argument concerns "the causal processes that theoretically transcend context" (p. 12). Idiographic-nomothetic synthesis counterfactuals are used by game theorists who "identify intriguing cross-case regularities" and "make predictions about how behavior will change as a lawful function of alterations in the probabilities or payoffs attached to courses of action" (p. 11).

Tetlock and Belkin (1996, p. 18) list six criteria for good counterfactual reasoning: (1) "*Clarity*: Specify and circumscribe the independent and dependent variables (the hypothesized antecedent and consequent); (2) *Logical consistency or cotenability*: Specify connecting principles that link the antecedent with the consequent and that are cotenable with each other and with the antecedent; (3) *Historical consistency (minimal-rewrite rule)*: Specify antecedent that require altering as few 'well-established' historical facts as possible; (4) *Theoretical consistency*: Articulate connecting principles that are consistent with 'well-established' theoretical generalizations relevant to the hypothesized antecdent-consequent link; (5) *Statistical consistency*: Articulate connecting principles that are consistent with 'well-established' statistical generalizations relevant to the antecedent-consequent link; (6) *Projectabilty*: Tease out testable implications of the connecting principles and determine whether those hypothe-

own and others' intellectual positions" (p. 13).

ses are consistent with additional real world observations." In general, I agree with this list. However, a key difference between the counterfactual analyses they are evaluating and my approach to counterfactual analysis is that they only explicitly consider one potential consequent: the analyst is either focused on arguing that no X would lead to no Y, or arguing that no X would have still lead to Y. I explicitly consider two potential consequents as possible: Y and no Y.

In concluding, Tetlock and Belkin (1996, p. 37) have a particularly evocative sentence: in cases where experimentation and statistical control are impossible, "The control groups exist - if indeed 'exist' is the right word - in the imaginations of political analysts who are left with the daunting task of reconstructing how history would have unfolded if causal variables of the past had taken on different values from the ones they actually did. The whole exercise starts to look hopelessly subjective, circular, and nonfalsifiable." My method is not a solution to *all* such woes. But for generalizable, narrowly-defined questions in cases where there is abundant historical evidence, my method does offer systematic, disciplined, transparent, and defensible approach to considering what would most likely have happened in a counterfactual world.

In their book on case studies and theory development, George and Bennett (2005, p. 167-168) have a brief discussion of counterfactuals in the context of controlled comparison: "Another way of attempting controlled comparison when two historical cases closely resembling each other cannot be located is to match the given case with an invented one that does. The case is, of course, a hypothetical one derived through counterfactual analysis of the existing case." George and Bennett have a list of concerns for what makes counterfactual cases difficult to wield for theory-testing. First, there must be good theoretical reason to suspect that the hypothesized X determines the Y, and second, there must be a good theory that supports how X causes Y. Third, counterfactual analysis is difficult when when there are multiple Xs that determine Y; fourth, counterfactual analysis cannot establish whether X was a cause of X if X is not a necessary cause of Y. Fifth, counterfactual

analysis does not work if the causes are interdependent and the X cannot be separated out; and sixth, counterfactual analysis cannot be sustained over a long chain of events (Ibid., p. 168-169). This is a good list of concerns to keep in mind when using counterfactual analysis for theory-testing, though the purpose of probabilistic counterfactual analysis is not theory-testing *directly*. Probabilistic counterfactual analysis plays a role in theory-testing by coding the dependent variable of the outcome that can then be considered in be considered in congruence testing (Ibid., Chapter 9) and surfacing intervening variables that can be examined in process-tracing (Ibid., Chapter 10).

In his review essay on counterfactuals, Levy (2008, p. 633) lists three criteria for evaluating counterfactual arguments: the clarity of the antecedents and consequents, the plausibility of the antecedent, and the conditional probability of the consequent given the antecedent. There is a trade-off when it comes to specifying the consequent. On the one hand, the consequent needs to be specific enough that the counterfactual argument is falsifiable, while on the other hand, if the consequent is too specific, then it is more likely to be simply false (p. 633-634). Similarly, there is a trade-off in constructing plausible antecedents, "between maximizing the plausibility of a counterfactual by minimizing the number of additional conditions necessary to sustain it, and selecting historically or theoretically meaningful counterfactuals" (p. 636). In other words, there is only a limited number of counterfactual worlds that can be brought about with a minimal rewrite of history, and the most theoretically interesting ones may not be amenable to being brought about by a minimal rewrite of history. And finally, whether the antecedent and necessary conditions to support the antecedent are likely to lead to the hypothesized consequent depends on the causal linkages that are "clearly specified, logically complete, and consistent wit the empirical evidence" (p. 638).[6] A good counterfactual argument should also meet the "projectability" criterion and the "proximity" criterion, be logically consistent, and especially aware of "redirecting counterfactuals" (which Lebow

[6]In an earlier article, Levy and Goertz (2007) lists the following criteria for evaluation of a counterfactual proposition: clarity, employing the minimal-rewrite rule, cotenability, consistency with well-established theoretical generalizations, historical accuracy, temporal proximity, and paying attention to redirecting counterfactuals.

calls "second-order counterfactuals").

Particularly relevant to my novel approach, Levy foreshadows the importance of probabilistic evaluation and the difficulty of sustaining counterfactual analysis over an excessively long causal chain. "It would also be inconsistent to try to support an argument that a particular outcome was contingent by demonstrating that a small change in the situation would have invariably led to a different outcome - one cannot support an argument for contingency by invoking a deterministic process. *To the extent possible, based on existing theory and empirical evidence, the author should give a sense of the likelihood that the consequent would have followed from the antecedent.* Counterfactual analysis, like any causal analysis, must *recognize that uncertainty about initial conditions and the pervasiveness of stochastic behavior limits even the best of our theories to relatively short-term predictions.* The longer the causal chain, the lower the probability of occurrence of the outcome at the end of the chain, particularly given the relative absence of sufficient conditions in social theory" (p. 642, emphasis added). Levy's general admonition to, *to the extent possible*, assign probabilities to the likelihood of different outcomes in a counterfactual world is what my novel method formalizes. [7]

In his book, Lebow (2010) identifies two distinct types of scholarly counterfactuals: "plausible" and "miracle" world counterfactuals. Plausible world counterfactuals "are intended to impress readers as realistic. ... realistic... in [the] sense of not violating our understanding of what was technologically, culturally, temporally or otherwise possible" (p. 44). Plausible worlds must not only be realistic, but they also "must have a significant probability of leading to the alternative outcome - and one we want to bring about" (p. 45). In contrast, miracle world counterfactuals "make inherently implausible changes in reality. They violate our understanding of what is 'realistic' or even conceivable, but they are valuable when they allow us to reason our way to the causes and contingency of real events, or the dynamics

[7] Levy also foreshadows my approach in a brief discussion about game theory: "Some theoretical approaches are more explicit than others about their counterfactual implications. A good example is game theory. A game tree specifies exactly what happens if actors make different choices, how other actors react, and the set of possible outcomes" (p. 630).

that govern them" (Ibid.). Some researchers only consider plausible counterfactual worlds legitimate, but "the utility of miracle counterfactuals does not depend on their realism but on the analytical utility of the alternative worlds they create" (p. 45-6).

Lebow (2010, p. 54-57) lists nine criteria for robust plausible counterfactual worlds: (1) realism, (2) clarity, (3) logical consistency or cotenability, (4) enabling counterfactuals do not undercut the antecedent, (5) historical consistency, (6) theoretical consistency, (7) avoidance of the conjunction fallacy, (8) recognition of the interconnectedness of causes and outcomes, and (9) consideration of second order counterfactuals. Pessimism in constructing a "robust" plausible counterfactual world - "a counterfactual whose antecedent we can *assert with confidence will result in the hypothesized consequent*" (p. 49, emphasis added) - is due to: "the statistical improbability of multistep counterfactuals, the interconnectedness of events, and the unpredictable effects of second-order counterfactuals" (p. 49). I will briefly summarize each challenge. This first challenges is the challenge of compound probability. It is challenging to start from a counterfactual antecedent and end at the hypothesized consequent because "The probability of a consequent is a multiple of the probability of each counterfactual linking it to the hypothesized antecedent" (p. 49). Say, for simplicity's sake, that the probability of each link in the causal chain from the antecedent to the consequent is 0.5. If there is only one link, the probability of the antecedent leading to the consequent is 0.5. If there are two links, the probability of the antecedent leading to the consequent is 0.5^2 or 0.25. If there are three links, the probability is 0.5^3 or 0.125. If there are five links, the probability is 0.5^5 or 0.031, and if there are eight links, the probability is 0.5^8 or 0.004, and so forth. The likelihood of a particular outcome resulting from a particular change becomes smaller and smaller the more steps there are in the causal chain from the counterfactual antecedent to the hypothesized consequent. Interconnectedness of events is the second challenge. "Surgical counterfactuals are no more 'realistic' than surgical air strikes. Causes are interdependent and have important interaction effects. Even plausible rewrites of history may alter the context in such a way as to render the consequent moot

or undercut the chain of events leading to it" (p. 50-1). Changing X may also change A, B, and C, which may or may not also have effects on Y. And the third and final challenge are second-order counterfactuals: "The problem of prediction is further complicated by the fact that history's clock does not stop if and when the hypothesized consequent is realized. Subsequent developments can return history to the course, or something close to it, from which the antecedent was intended to divert it" (p. 51). Changing X at one point in history does not guarantee that X will not occur at a later point in the counterfactual history. Antecedents that were thought to have major effects on the consequent may end up having only small effects on the consequent in the long-run. For these three reasons, "Plausible rewrite counterfactuals are easy to create but extraordinarily difficult, if not impossible, to control" (Lebow 2010, p. 51).

While acknowledging that plausible counterfactual worlds can be difficult or impossible to control, Lebow (2010) notes that counterfactuals can be useful even if they cannot be strictly controlled. "Counterfactuals can help ... [make] scholars more sensitive to contingency, [help] them work through the implications of existing theories and identify gaps and inconsistencies in them. ... Counterfactual experimentation ... can be used effectively *without the same degree of confidence that specific antecedents will lead to specific consequents.* ... We should *worry less about the uncertainty of counterfactual experiments and profit more from their mind-opening implications*" (p. 53-4, emphasis added).

The main advantage of a counterfactual thought experiment over a real world cross-case or intra-case comparison is that we can be much more precise in changing *just* the variable we are interested in observing the effects of. The main disadvantage of the counterfactual thought experiment is that it is dealing with an unobservable world. There is no real world evidence to directly observe. What would happen in the counterfactual world must be extrapolated from evidence from the real world. This limits what questions counterfactual thought experiments can answer to questions where there is sufficient evidence from the real world to reasonably support extrapolation. However, the lack of data is not a problem

unique to counterfactual thought experiments. Even when studying the real world, data is limited and biased. Interviewees may not remember the past so clearly or deliberately paint themselves in a better light. Government classification schemes may prevent critical information for an argument from being released to scholars. Incriminating evidence is also oftentimes outright destroyed. We account for gaps in evidence in any scholarly endeavor. Though extrapolation is baked into the counterfactual method, extrapolation is an inherent part of any academic analysis that is more than descriptive observations of the real world. "The inherently speculative nature of counterfactuals make many scholars, especially historians, wary of them. We should nevertheless recognize that counterfactual experiments can sometimes be conducted in an evidence-rich environment" (Lebow 2010, p. 30).

Most existing works on counterfactual analysis have lists of what makes a counterfactual analysis a good counterfactual analysis - whether of the plausible or the miracle kind. The most innovative breakthrough in the conduct of counterfactual analysis is a study of the causes of the Iraq War by Harvey (2012). Harvey develops comparative counterfactual analysis which is based on the premise that "at least two mutually exclusive counterfactual arguments emerge from every thesis" (p. 27). Harvey's strength is in making explicit what has always been implicit. Like my own approach, he is not just focused on whether no X leads to either Y or no Y, but he is equally focused on how no X may lead to Y as well as how no X may lead to no Y. He continues, "The evidence for and against both counterfactuals must be discussed together, because the strengths of one are directly relevant to uncovering weaknesses of the other and vice versa" (p. 27). In Table 1, I show a generic adaptation of his comparative approach to counterfactual analysis (p. 29). In Harvey's approach to counterfactual analysis, analysts must pay equal attention to evidence that falls in the top left (No X, No Y) and bottom left (No X, Y) cells of the table.[8] Comparative counterfactual

[8]Harvey critiques discussions about the causes of the Iraq War by saying that analysts focus too much on collecting data that supports the Gore (No X)-peace (No Y) argument, while ignoring and not even examining the data that supports the Gore (No X)-war (Y) argument. Evidence that supports the Gore-peace argument would fall in the top left cell, and evidence that supports the Gore-war argument would fall in the bottom left cell. Harvey's innovation is to be explicit that counterfactual analysis requires collecting data on the likelihood of either option, or at the very least, explicitly thinking about the flip side of the counterfactual

Table 1: Comparative Counterfactual Analysis (CCA)

		Independent Variable	
		No X (Counterfactual Premise)	X (Real World Premise)
Dependent Variable	No Y	Does extrapolating evidence from real world support this conclusion?	N/A
	Y	Does extrapolating evidence from real world support this conclusion?	Real world

analysis is a contribution to the disciplined and systematic study of counterfactuals. Probabilistic counterfactual analysis is another step forward in disciplining and systematizing the analysis of counterfactual worlds.

Cotenability: Limitations on Counterfactual Analysis

The kinds of questions counterfactual analysis is suited to address is limited by the requirements of cotenability. Formally, cotenability is the requirement that (1) generally accepted theories, appropriate assumptions, sound logic, and relevant historical evidence about the initial conditions when the manipulation is made imply that the outcome flows from the initial conditions with the manipulation, and (2) *that the initial conditions that implied the outcome could have existed in much the same way even with the manipulation* (Goodman 1947; Goodman 1983). Intuitively and practically, cotenability is about how similar the real and counterfactual worlds look. Is the manipulation small or insignificant (but interesting) enough that the real world still provides sufficient empirical grounds on which to predict what could likely happen in the counterfactual world? Or is the manipulation so large or significant that the real world no longer provides any empirical guidance on what could happen in the counterfactual world?

argument one favors. For the Bush (Y)-war (X) argument to be correct, i.e. for Bush to have been the cause of the Iraq War, the evidence for Gore-peace counterfactual must be greater than the evidence for Gore-war counterfactual. After examining Al Gore's values, policy preferences and misperceptions (Chapter 2), Gore's advisers' values, policy preferences and misperceptions (Chapter 3), domestic and congressional politics and political support (Chapter 4), organizational routines, American intelligence failures and miscalculations (Chapter 5), society, the media and public opinion (Chapter 6), international politics, global WMD consensus and UN balancing (Chapter 7), and Saddam's intelligence failures and miscalculations (Chapter 8), Harvey himself concludes that the evidence for Gore-war outweighs the evidence for Gore-peace.

In counterfactual analysis, the ideal is to remove X and have a counterfactual world that is otherwise identical to the real world. However, the ideal is impossible, and the larger and more significant the X, the less similar the counterfactual world will be to the real world. If X is too broad, e.g. nationalism, imperialism, militarism, then removing X may change the counterfactual world too much from the real world to make reasonable extrapolations from real world evidence. A given manipulation may be so extreme that initial conditions would no longer exist in the counterfactual world sufficiently similar to initial conditions in the real world. The requirement of cotenability means that only some questions are amenable to counterfactual analysis.

Fearon (1991, p. 193-4) expects that, "in practice, the cotenability requirement will be more plausibly satisfied for small causes, such as specific policy decisions, than for big causes, such as nationalism, imperialism, or a cult of the offensive. History often provides evidence that leaders considered several possible choices at certain junctures, and in some instances it may be feasible to imagine a different choice without changing other major influences on the outcome in question. The fewer the changes from the actual world required by a counterfactual supposition, the easier it will be to draw and support causal inferences, and the more defensible they will be." Questions about the role of leaders falls somewhere in the middle of this range. A counterfactual world where a particular leader is different is more cotenable with the real world than a counterfactual world where nationalism, imperialism, or a cult of the offensive is removed. However, a counterfactual world where a particular leader is different is less cotenable with the real world than a counterfactual world where the same leader was in power but merely made a different policy decision.[9]

[9]There are cases where leaders could have been different for very historically plausible reasons. An example would be Winston Churchill replacing Neville Chamberlain instead of Halifax in May 1940 (Khong 1996, p. 113-114).

Figure 1: Probabilistic Approach to Counterfactual Analysis

$$\text{Observed World: } X = 1$$
$$P(Y = 1|X = 1) = 1$$
$$P(Y = 0|X = 1) = 0$$
$$P(Y = 1|X = 1) + P(Y = 0|X = 1) = 1$$

$$\text{Counterfactual World: } X = 0$$
$$0 < P(Y = 1|X = 0) < 1$$
$$0 < P(Y = 0|X = 0) < 1$$
$$P(Y = 1|X = 0) + P(Y = 0|X = 0) = 1$$

Probabilistic Counterfactual Analysis

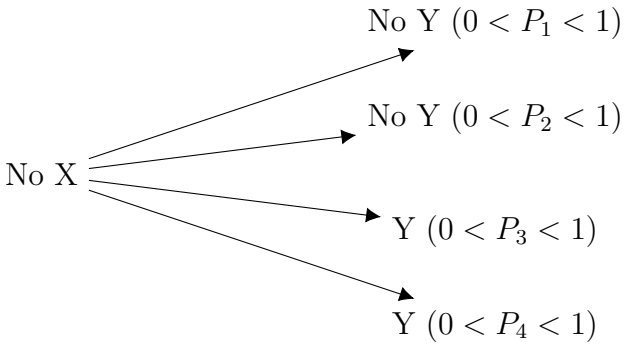
My approach, instead of arguing that removing X *deterministically* leads to the presence or absence of Y, *probabilistically* evaluates whether the presence or absence of Y is more likely. Instead of an all-or-nothing approach to the presence or absence of Y in the absence of X, I evaluate the relative probabilities that Y would still have happened even if there was no X, i.e. $P(Y = 1|X = 0)$, and that Y would not have happened if there was no X, i.e. $P(Y = 0|X = 0)$. The counterfactual is limited so that the only outcomes we are interested in is Y or no Y, and defined so that the probability of the two potential outcomes adds up to 1. See Figure 1 for mathematical notation.

My probabilistic approach shifts the analytical focus from the *one* counterfactual *world* that *must* follow after removing X to the *many* different counterfactual *paths* that *could* ensue after removing X. These paths, all starting from no X, can lead to either no Y or to Y. The probability of Y is found by adding all the paths that led to Y, and the probability of no Y is found by adding the probabilities of all the paths that led to no Y. Thinking of the counterfactual analysis in terms of many different paths sensitizes us to carefully and systematically think through many different possibilities instead of quickly and narrowly trying to focus on the one "right" answer about what would happen in the counterfactual world. See Figure 2 for a visualization.

Figure 2: Paths in Observed and Counterfactual Worlds

$$P_0 = 1$$
$$P_0 = P(Y = 1|X = 1)$$
$$0 = P(Y = 0|X = 1)$$

(a) One Path in the Observed World



$$P_1 + P_2 + P_3 + P_4 = 1$$
$$P_1 + P_2 = P(Y = 0|X = 0)$$
$$P_3 + P_4 = P(Y = 1|X = 0)$$

(b) Multiple Paths in the Counterfactual World

Boundedly Arbitrary Rewrites

What kind of rewrites are permissible in the probabilistic approach to counterfactual analysis? I suggest taking the most useful elements from both plausible and miraculous rewrites: what I call the boundedly arbitrary rewrite. Plausible rewrites must be realistic in that only small changes are needed to change the antecedent. Examples include a barely-successful assassination in the real world failing in the counterfactual world, or an extremely close election going to the other candidate instead in the counterfactual world. However, that severely limits the range of counterfactual worlds that can be considered. Miraculous rewrites require larger changes and may be less realistic, but "The value of [miraculous] counterfactuals is based entirely on their ability to provoke and inspire and compel researchers to think about issues and problems they would not otherwise recognize or address, or to look at them in a new light. For a field where careful technical work is increasingly valued over imagination, miracle-world counterfactuals have the potential to refocus our attention on a deeper set of important questions" (Lebow 2010, p. 58). Examples of miraculous rewrites include a young, healthy person dying inexplicably in the counterfactual world, or a landslide election result being thrown out in the counterfactual world.

I argue that an *arbitrary* rewrite is acceptable if it helps answer a question we are interested in.[10] Fearon (1996, p. 61) would also accept the suitability of what he calls miracle causes in my approach, because I am "treating the specific event to be explained as an instance of a class of events." If the purpose of the counterfactual thought experiment is to determine the effect of X on Y, and the presence or absence of Y is a general phenomenon, then even if removing X is arbitrary, it is an acceptable change given the greater analytical purpose. For example, if we want to estimate the effect of a particular leader on the formation of a given alliance, then even if removing the particular leader from power is arbitrary, such a counterfactual thought experiment is still the best way to isolate the effect of the

[10] I use the word arbitrary in the sense that it will be historically arbitrary for this change. However, it is a targeted, bounded, and *non*-arbitrary rewrite in the analytical sense, as the analyst is executing this rewrite to answer a specific, generalizable question that they have.

particular leader on the given alliance.

However, the arbitrariness of the change must be *bounded* for two reasons. First, the counterfactual world still has to be similar enough to the real world that evidence from the real world can be used to extrapolate which counterfactual paths are more likely, i.e. it must still meet the cotenability requirement. Second, if too many variables are changed, it defeats the purpose of trying to isolate a single hypothetical cause through the use of counterfactual thought experiments. Turning to Fearon (1996, p. 64-65) again, "Regarding the ... problem of how we should imagine the miracle cause being inserted or subtracted in the counterfactual scenario, I can only suggest a Lewislike closeness criterion: introduce the miracle by making as few changes as one can in the actual world." If a change is too extremely arbitrary, it no longer useful. To continue with the example of trying to determine the effect of a particular leader: after arbitrarily removing the particular leader, this particular leader should not be replaced by a leader from the opposition party if the opposition party was so minor that they had no realistic chance of seizing political power or a leader who would have been institutionally barred from holding that position. Instead this particular leader should be replaced by someone who had a realistic chance of holding that office - such as the leader from the opposition party if the opposition party was on equal terms with the real world leader's party, or whichever leader was institutionally next in line to assume the mantle. Predecessors, successors, and second-in-commands are all viable alternative leaders in a boundedly arbitrary rewrite. With predecessors and successors, there is historical evidence that there was a viable path for them to attain the office. With the second-in-command, it is institutionally viable for them to have attained the office even if they may not have actually attained the office in the real world.[11]

[11]In particular, the advantages of analyzing a counterfactual world in which the predecessor remained in power longer than they did in the real world is that they likely left a record of their opinions about future policy debates and may have continued to participate in policy debates even after they left office. The disadvantage of analyzing a counterfactual world in which the successor came to power earlier that they did in the real world is that whatever they said and did in the real world with regards to the policy outcome of interest is likely to be biased by the fact that the policy is already enacted. If the policy was popular, then they are likely to say they supported the policy, but it may be difficult to extrapolate backwards with confidence that they would have supported it before it was enacted. If the policy was unpopular, then they

To recap, a boundedly arbitrary rewrite combines the most useful elements of plausible and miraculous rewrites. Because the historical manipulation is arbitrary in order to answer a question we are interested in, it resembles a miraculous rewrite. However, because the arbitrariness of the manipulation is bounded by cotenability concerns, it resembles a plausible rewrite. A boundedly arbitrary rewrite consists of *a miraculous rewrite brought about by the* most *plausible rewriting possible.*

The Five-Step Process

The process of applying probabilistic counterfactual analysis must be tailored to each question. However, for any application there is a generic five-step process that can be followed. *First*, observe the real world to confirm X appears before Y. *Second*, create a counterfactual world in which X is absent. X may be removed through a boundedly arbitrary rewrite - a miraculous rewrite brought about by the *most* plausible rewriting possible. *Third*, construct a decision tree by identifying decision points. The decision tree starts with no X, and has some paths that lead to no Y and some paths that lead to Y. *Fourth*, convert the decision tree into a probability tree to calculate how confident we would need to be about choices made at each decision point to conclude that Y was more likely under counterfactual premises or to conclude that no Y was more likely under counterfactual premises. *Fifth*, weigh the historical evidence to make an assessment about whether Y or no Y was more likely.

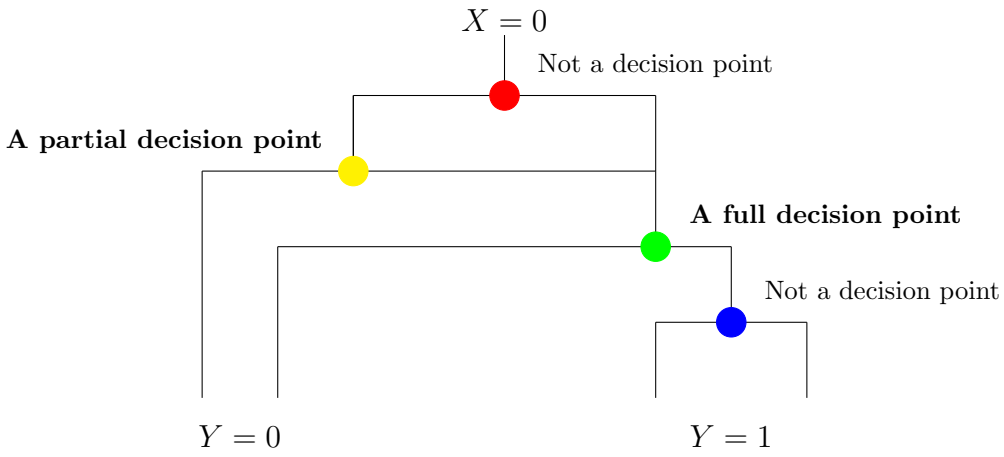
Decision Trees: Identifying Decision Points

The best way to visualize counterfactual paths in a counterfactual world is through decision trees. In constructing a decision tree, each node is a decision point. A *decision point* is a choice that rules out one of the two potential outcomes, Y or no Y, thereafter.[12] To

are likely to say they opposed the policy, but again, this may not accurately reflect what they believed at the time of policymaking. It is possible, if the successor was involved in policy debates before they assumed office, to find records of what they believed un-biased by the outcome of the real world.

[12]Decision points functionally are the same as critical junctures in the study of historical institutionalism insofar as both decision points and critical junctures close off alternative options later down the road. I choose to use a different term to distinguish it from critical junctures in historical institutionalism, because the changes I focus on are more micro-level, and the period of equilibrium or stasis between each decision point

Figure 3: Identifying Decision Points

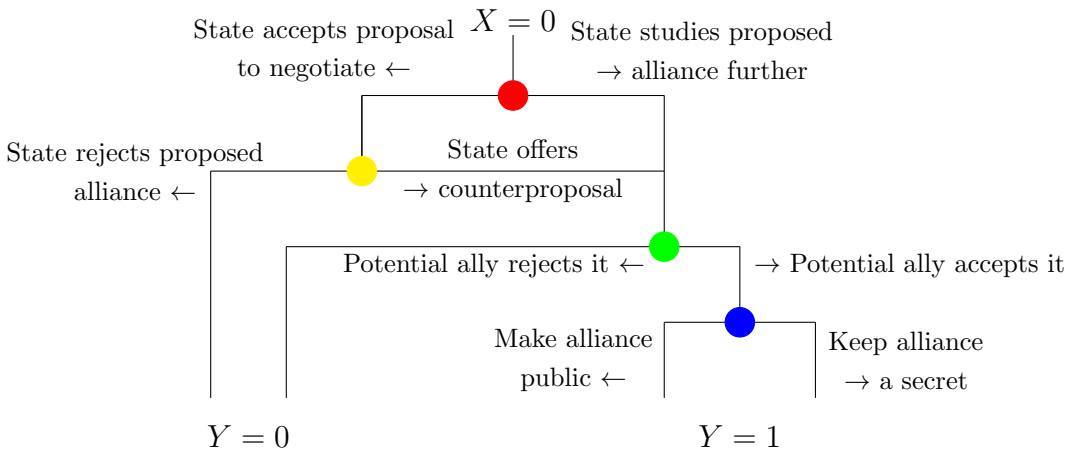


illustrate, take a look at Figure 3. Starting from the top of the figure at $X = 0$ and moving to the bottom of the figure at either $Y = 0$ or $Y = 1$, the red, yellow, green, and blue dots all represent choices that an actor could make. However, only the yellow and green dots represent a choice that is a decision point. The choice made at the red dot is not a decision point because whether you go right or left at the red dot, you can still get to either $Y = 0$ or to $Y = 1$ eventually. The choice made at the blue dot is not a decision point because whether you go right or left at the blue dot, you can only end up at $Y = 1$. The choice made at the yellow dot is a *partial decision point*, because if you go left at that point you can only get to $Y = 0$ and are ruling out $Y = 1$ as a potential outcome thereafter. If you go right at the yellow dot, both $Y = 0$ and $Y = 1$ are still potential outcomes. The choice made at the green dots is a *full decision point*, because if you go left at the green dot, you can only get to $Y = 0$ thereafter, and if you go right at the green dot, you can only get to $Y = 1$ thereafter. Though the real world is complicated, I suggest simplifying the choice faced at each decision point to two options. These two options must be constructed to be exhaustive and mutually exclusive.

I will give concrete examples in the context of alliance formation to draw the distinction between choices that are decision points and choices that are not decision points as illustrated

is much shorter or non-existent. For more on critical junctures in historical institutionalism, see Capoccia (2016); Capoccia and Kelemen (2007); Collier and Munck (2017); Collier and Collier (1991); Thelen (1999).

Figure 4: Example of Decision Tree for Alliance Formation



in Figure 4. A non-decision point choice like the red dot that does not close the path to either $Y = 0$ or to $Y = 1$ could be the choice to either immediately accept a potential ally's overtures to negotiate or the choice to study the option further before accepting a potential ally's overtures. Whether the potential ally's overtures were immediately accepted or merely given further consideration, alliance ($Y = 1$) or no alliance ($Y = 0$) are both still potential outcomes down the road. In this case, say that immediate acceptance and beginning of negotiations is going left to the yellow dot and further study of the option is going right to the green dot. After immediately accepting the potential ally's overtures and going left at the red dot, a partial decision point like the yellow dot could be the state's choice to reject the proposed alliance or the choice to make a counterproposal. If the state rejects the proposed alliance, going left at the yellow dot, this choice rules out alliance ($Y = 1$) as a potential outcome. Only no alliance ($Y = 0$) is possible after this choice is made. However, if the state offers a counterproposal, both alliance ($Y = 1$) and no alliance ($Y = 0$) are still possible depending on the potential ally's reaction.

The green dot can be reached by one of two ways: either the state had immediately begun negotiations with the potential ally (going left at the red dot) and decided to make a counterproposal (going right at the yellow dot) or the state had considered the matter further and decided to accept the potential ally's overtures (going right at the red dot).

The green dot in this case could be the potential ally's decision to either accept the state's delayed response or counterproposal. The green dot is a full decision point, because if the potential ally rejects the state's delayed response or counterproposal, going left at the green dot, alliance ($Y = 1$) is no longer a potential outcome, only no alliance ($Y = 0$) is, and if the potential ally accepts the state's delayed response or counterproposal, going right at the green dot, no alliance ($Y = 0$) is no longer a potential outcome, only alliance ($Y = 1$) is. A non-decision point like the blue dot would be a choice about which portions of the alliance would be made public and which portions would be kept secret. Unless the state or its potential ally cared about the publicity or the secrecy of the alliance so much that they were willing to walk away from the negotiated alliance without signing it, this decision should not change the outcome of alliance formation ($Y = 1$). This choice does not change what the potential outcome would be at this point.

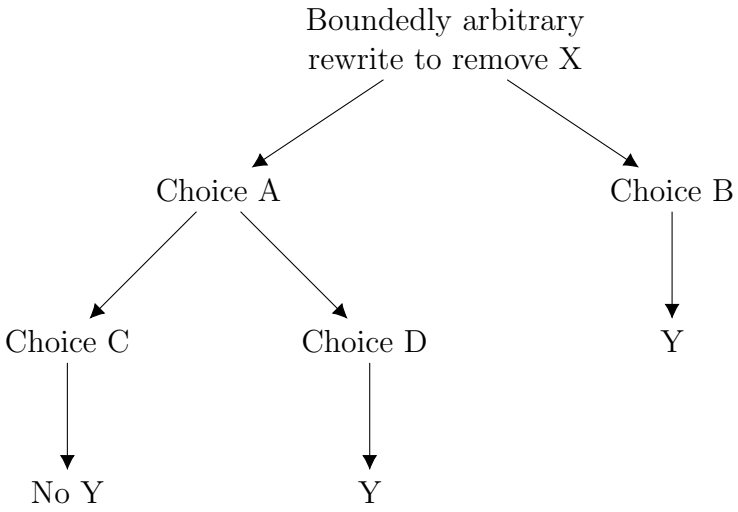
Including all four choices that get made at the red, yellow, green, and blue dots in a decision tree would make the analysis more realistic. However, only the choices that get made at the yellow and green dots are decision points, and analysts should devote most of their time analyzing whether the actor would more likely go right or left at these decision points. Analysts should focus on substantively important decision points rather than trying to document every superficial choice that would not necessarily change the outcome we are ultimately interested in: Y or no Y. Decision trees are not better when they are longer. They are better when they help analysts identify where in the decision-making process they should focus on in thinking about counterfactual paths that could lead to the different potential outcomes in the counterfactual world. These decision points are identified on a case-by-case basis based on a close reading of how events unfolded in the real world. Two steps is a minimum for modeling interactions like alliance negotiations, where both parties to the proposed alliance have a say. Occasionally the decision tree may have more than two steps, but only if events in the real world strongly suggest that additional decision points need to be considered. This is in adherence to the proximity criterion laid out in Fearon (1996, p. 66):

"Consider only thought experiments in which the hypothetical antecedent and consequent are close together in time and are separated by a small number of causal steps."

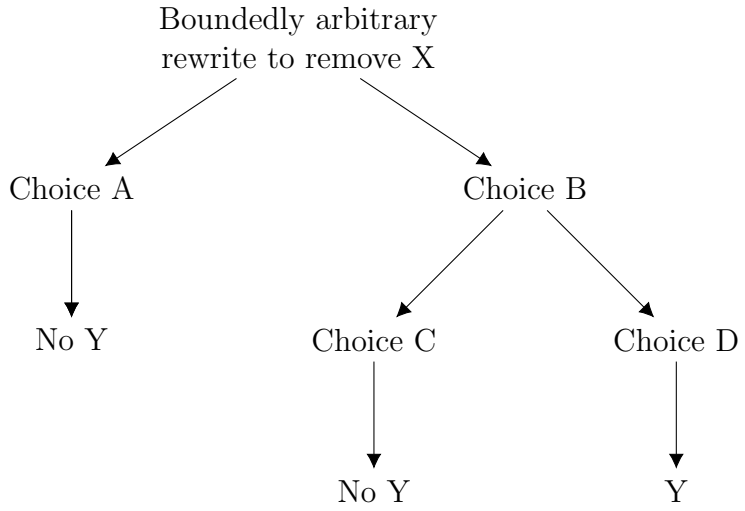
Figures 5a and 5b are generic two-step decision trees. As you can see, in Figure 5a the choice between A and B is a partial decision point, because if the first-moving actor chooses B, it rules out the possibility of No Y as an outcome thereafter. In Figure 5b, the choice between A and B is also a partial decision point, because if the first-moving actor chooses A, it rules out the possibility of Y thereafter. In both Figures 5a and 5b, the choice between C and D is a full decision point, because if the responding actor chooses C, then the outcome will be No Y, and if the responding actor chooses D, then the outcome will be Y. Figures 6a and 6b are generic three-step decision trees. Whether a particular case requires two or three steps should be determined on a case-by-case basis, based on how events actually unfolded. The *key* to probabilistic counterfactual analysis is that the probabilities for the two options at each node must add up to 1. This is done by ensuring that the two options at each decision point are exhaustive and mutually exclusive.

A concept that I will become more relevant in the next section that I will explain now is the *improbable path*. Simply, this is the longest path. It is the path that leads to Y when all the other branching paths lead to no Y, or the path that leads to no Y when all the other branching paths lead to Y. The longest path, by virtue of the way this method focuses on decision points, is always the *only* path that leads to either no Y or Y. All the other paths will lead to the other potential outcome. Depending on the historical evidence, this is *not the historically most improbable path*, but it is *mathematically the most improbable path* because of the number of times the probabilities must be multiplied along the path to get to the end of the improbable path (whether it is Y or no Y) is greater than the number of probabilities that must be multiplied to get to the end of any of the branching paths. Furthermore, because the probabilities of the potential outcome at the end of each of the branching paths get added together, it makes the potential outcome at the end of the improbable path the least likely - again, from a purely mathematical and not historical perspective. Below, I have

Figure 5: Generic Two-Step Decision Trees

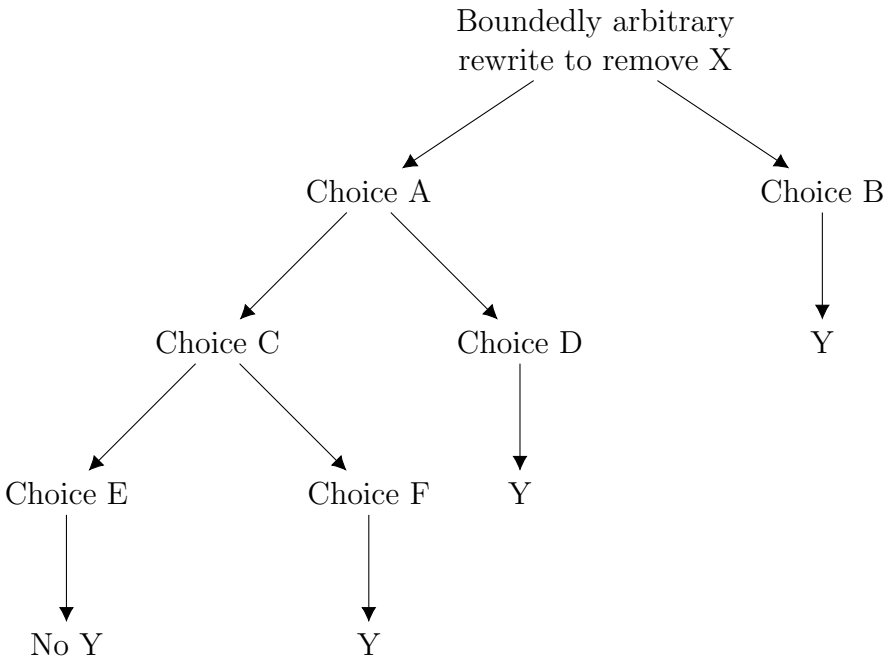


(a) Generic Two-Step Decision Tree 1

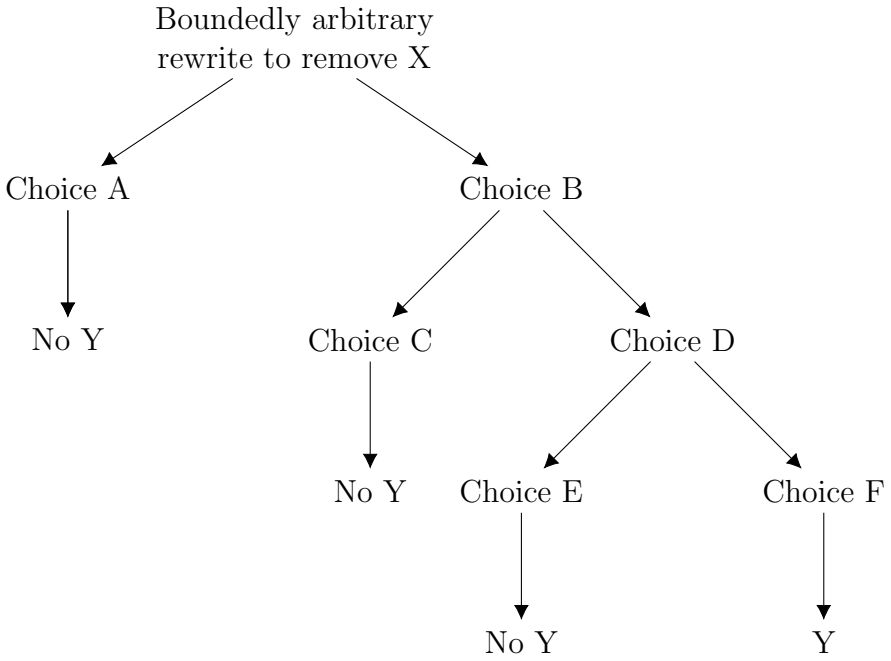


(b) Generic Two-Step Decision Tree 2

Figure 6: Generic Three-Step Decision Trees



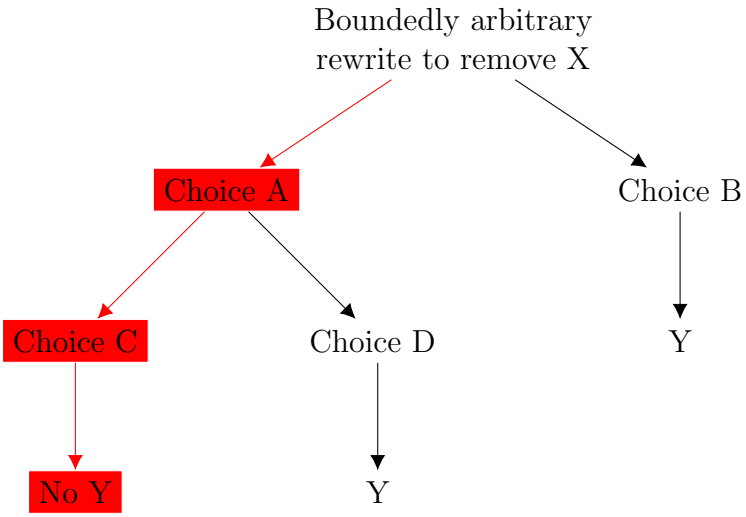
(a) Generic Three-Step Decision Tree 1



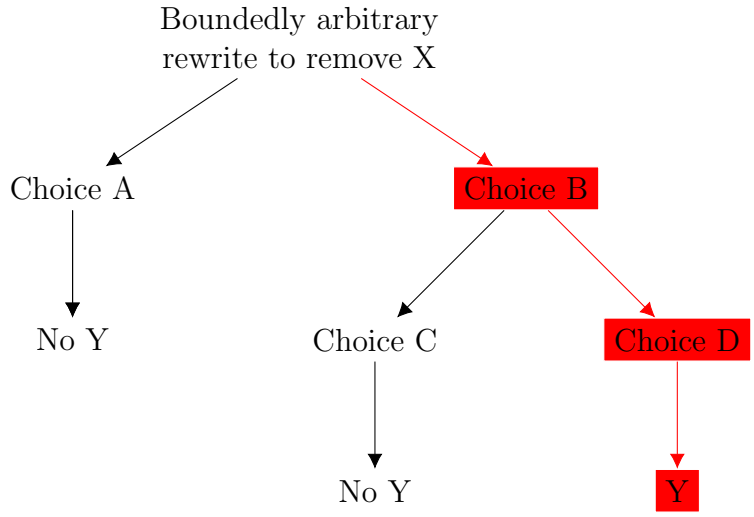
(b) Generic Three-Step Decision Tree 2

all of the generic decision trees again, this time with the improbable path highlighted in red.

Figure 7: The Improbable Path in Generic Two-Step Decision Trees

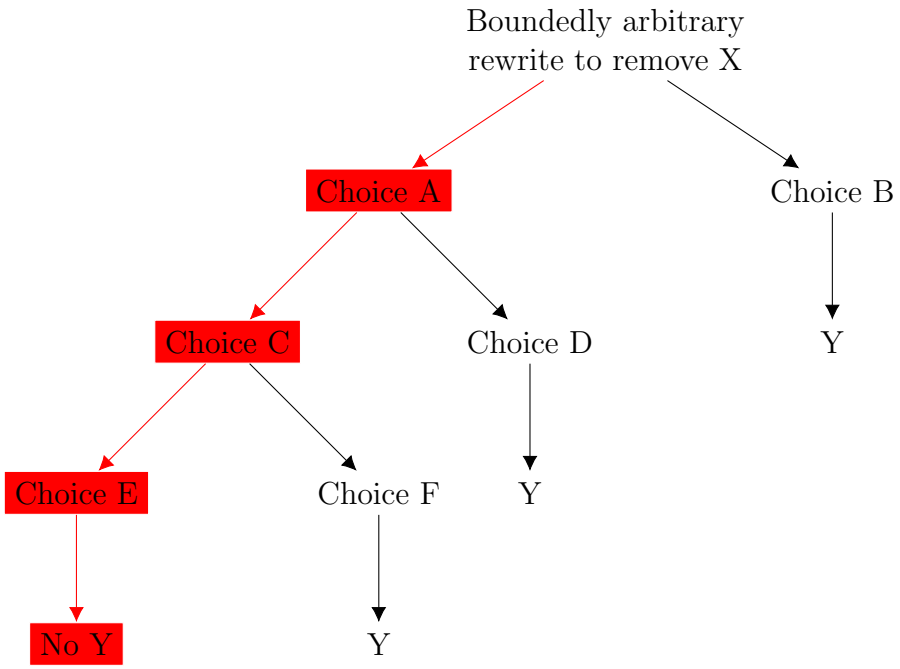


(a) The Improbable Path in Generic Two-Step Decision Tree 1

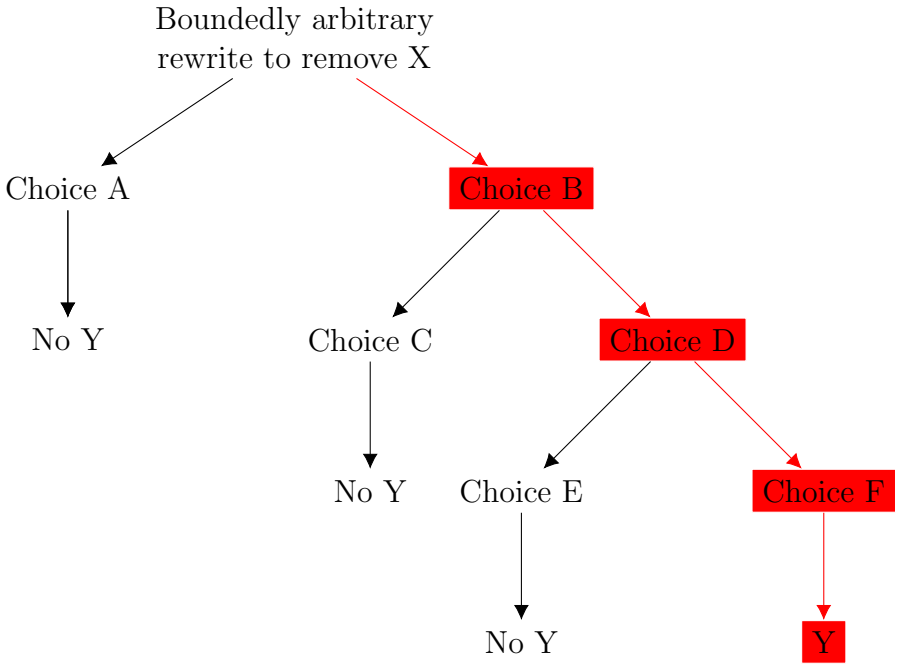


(b) The Improbable Path in Generic Two-Step Decision Tree 2

Figure 8: The Improbable Path in Generic Three-Step Decision Trees



(a) Generic Three-Step Decision Tree 1



(b) Generic Three-Step Decision Tree 2

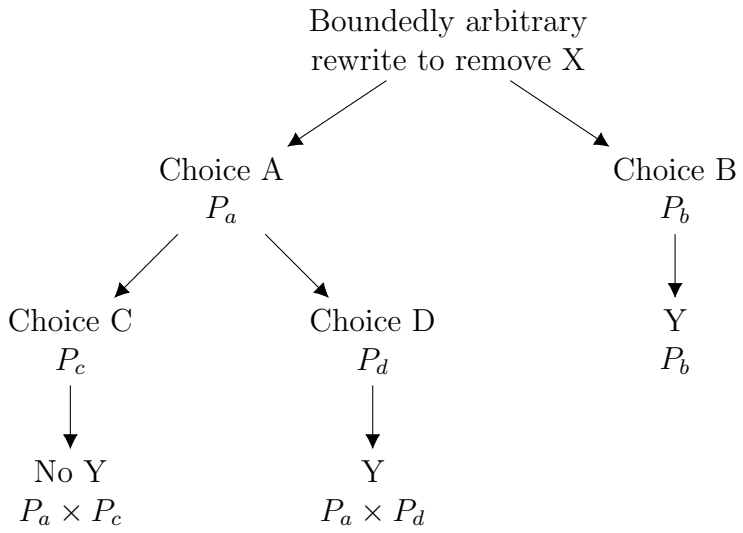
Probability Trees and Historical Evidence: Calculating Minimum Confidence and Assessing Relative Probabilities

The next step after constructing the decision trees is to convert them into probability trees. By doing so, we can assess how confident we need to be in the historical evidence to reach conclusions about which potential outcome was more likely in this counterfactual world. Figures 9a and 9b are generic two-step probability trees, and Figures 10 and 11 are generic three-step probabilities. Because the probability of the two options at each node always add up to 1, the probability of each particular path is simply a multiplication of the probabilities of all the choices made along that path. The probability of outcome No Y can be found by adding the probabilities of each path that led to No Y, and the probability of outcome Y is found by adding the probabilities of each path that led to Y.

One outcome is more likely than the other if the probability of that outcome is greater than 0.50. In other words, we can assess whether Y or no Y is more likely by assessing whether the probability of reaching the end of the improbable path is greater than or less than 0.50. If the improbable path of the decision tree ends in no Y, as Figures 7a and 8a do, and the probability of reaching the end of that path is greater than 0.50, then no Y is more likely than Y in this counterfactual world and vice versa. If the improbable path of the decision tree ends in Y, as Figures 7b and 8b do, and the probability of reaching the end of that path is greater than 0.50, then Y is more likely than no Y in this world and vice versa.

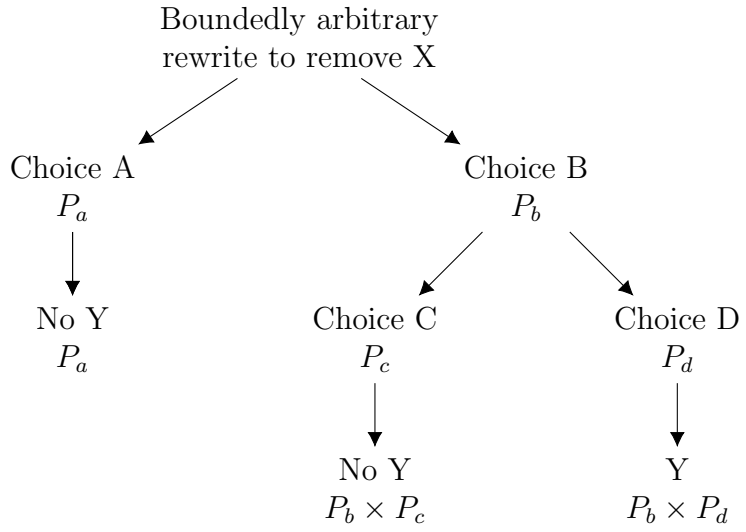
To make this assessment, the method takes two approaches based on the quality and abundance of historical evidence relevant to the decision point. If the historical evidence from which to extrapolate that one option would have been chosen over the other is ambiguous or limited, then we go ahead and assign the probability of 0.50 to both options for that decision point. If there is historical evidence from which to extrapolate that one option would have been chosen over the other, then the probability tree tells us *how* confident we have to be to reach a conclusion that one option would have been chosen over the other at each decision point, and that one potential outcome was more likely than the other under

Figure 9: Generic Two-Step Probability Trees



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_a \times P_c = P(Y = 0|X = 0)$$
$$P_b + (P_a \times P_d) = P(Y = 1|X = 0)$$

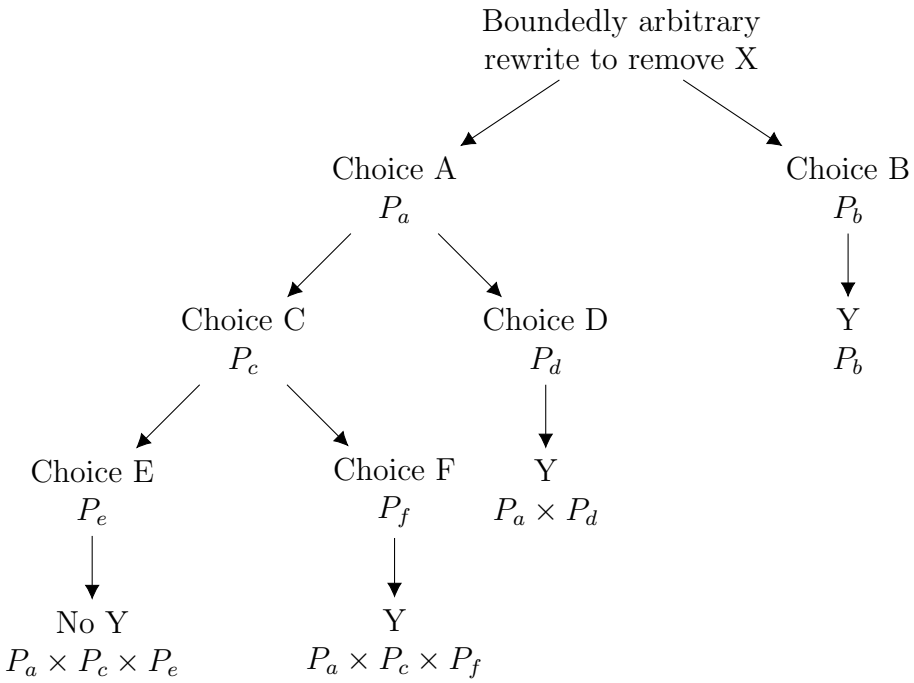
(a) Generic Two-Step Probability Tree 1



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_a + (P_b \times P_c) = P(Y = 0|X = 0)$$
$$P_b \times P_d = P(Y = 1|X = 0)$$

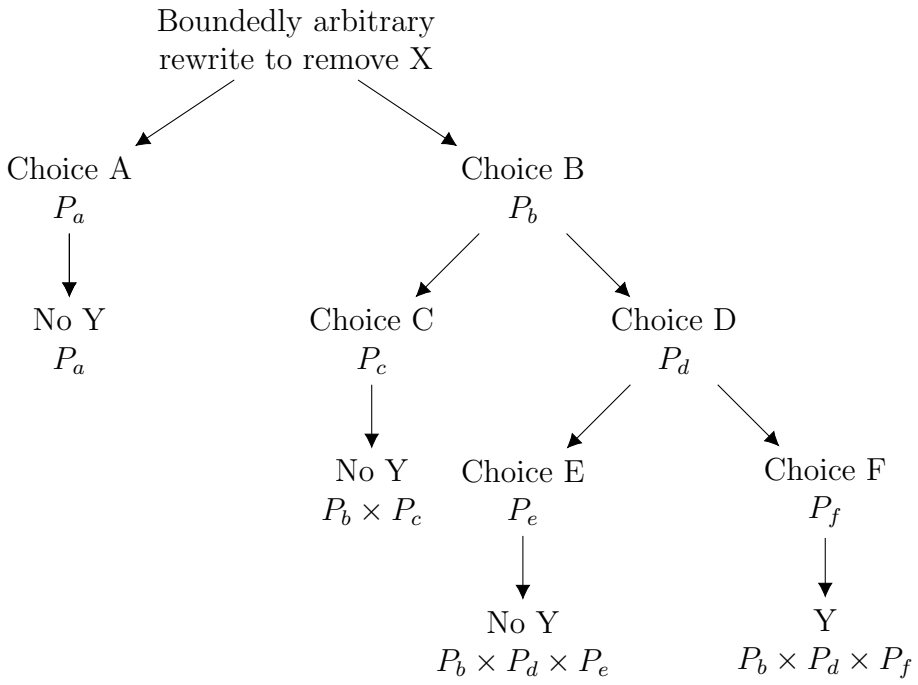
(b) Generic Two-Step Probability Tree 2

Figure 10: Generic Three-Step Probability Tree 1



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_e + P_f = 1$$
$$P_a \times P_c \times P_e = P(Y = 0|X = 0)$$
$$P_b + (P_a \times P_d) + (P_a \times P_c \times P_f) = P(Y = 1|X = 0)$$

Figure 11: Generic Three-Step Probability Tree 2



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_e + P_f = 1$$
$$P_a + (P_b \times P_c) + (P_b \times P_d \times P_e) = P(Y = 0|X = 0)$$
$$P_b \times P_d \times P_f = P(Y = 1|X = 0)$$

this counterfactual premise.

Decision Points for which there is Ambiguous or Limited Historical Evidence

Evidence is *ambiguous* when the actor confronted a very similar choice or dilemma in the real world, but this choice or dilemma is sufficiently different in the counterfactual world from the real world that we cannot simply extrapolate from their real world behavior to their most likely counterfactual world behavior. Historical evidence is most likely to be ambiguous if the boundedly arbitrary rewrite changed the world in a way that makes the actor's reasoning in the real world not applicable in the counterfactual world.

Evidence is *limited* when the actor did not confront a similar choice or dilemma, or if they did, we do not have documented evidence on how they handled the situation. We do not have real world evidence from which to extrapolate their counterfactual world behavior.

When the historical evidence is so ambiguous or limited that it does not provide any support for extrapolating that one option was more likely to be chosen than the other option at this decision point, then the analyst assigns the probabilities as 0.50 for both options at that decision point. Giving both options a 50% chance is statistically saying historical evidence cannot provide any support to extrapolate which option was more or less likely to be chosen in the counterfactual world. What could happen in the counterfactual world is a coin-toss.

Decision Points for which there is Clear and Abundant Historical Evidence

Evidence is *clear and abundant* when the actor confronted a very similar choice or dilemma in the real world, this choise or dilemma is sufficiently similar in the counterfactual world that we can extrapolate from real world behavior, and we have documented evidence about how they handled the situation and why they handled the situation the way that they did. This documented evidence allows us to extrapolate with a high degree of confidence how the actor would behave in the counterfactual world. But the question still remains: when there is clear and abundant historical evidence that one option was more likely to be chosen than

the other option, *how* confident do we need to be?

In a two-step probability tree in which the improbable path leads to no Y (Figure 7a), we need to be at least 71% confident that at each decision point, the option that leads to no Y will be chosen. The 71% number comes from: $\sqrt{0.50} \approx 0.7072$. To put it another way, to conclude that $P(Y = 0|X = 0) > 0.50$ in a counterfactual world set up as Figure 7a, it must be true that $P_a > 0.7072$ *and* $P_b > 0.7072$.[13] To be 71% confident that the relevant actor would make Choice A is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite *but is otherwise exactly the same as the real world*) 10,000 times over, in at least 7,072 of these worlds, the actor would make Choice A. To be 71% confident that the relevant actor would make Choice C is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choice A has already been made) 10,000 times over, in at least 7,072 of these worlds, the actor would make Choice C. See Figure 12a for a visual of an example where no Y is more likely than Y.

In a two-step probability tree in which the improbable path leads to Y (Figure 7b), we need to be at least 71% confident that at each decision point, the option that leads to Y will be chosen. To put it another way, to conclude that $P(Y = 1|X = 0) > 0.50$ in a counterfactual world set up as Figure 7b, it must be true that $P_b > 0.7072$ *and* $P_d > 0.7072$.[14] To be 71% confident that the relevant actor would make Choice B is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite *but is otherwise exactly the same as the real world*) 10,000 times over, in at least 7,072 of these worlds, the actor would make Choice B. To be 71% confident that the relevant actor would make Choice D is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choice B has already been made) 10,000

[13]Technically, there is a number of different combinations of probabilities that could lead to the outcome at the end of the improbable path being greater than 0.50. For example, even if we were only 62.5% confident that one of the decision points would lead to no Y, if we are more than 80% confident that the other decision point would lead to no Y, the probability of no Y is still 0.50. However, for simplicity's sake, I focus on being at least 71% confident at each decision point.

[14]Same as the footnote above.

times over, in at least 7,072 of these worlds, the actor would make Choice D. See Figure 12b for a visual of an example where Y is more likely than no Y.

In a three-step probability tree in which the improbable path leads to no Y (Figure 8a), we need to be at least 80% confident that at each decision point, the option that leads to no Y will be chosen. The 80% number comes from $\sqrt[3]{0.50} \approx 0.7938$. To put it another way, to conclude that $P(Y = 0|X = 0) > 0.50$ in a counterfactual world set up as Figure 8a, it must be true that $P_a > 0.7938$, $P_c > 0.7938$, *and* $P_e > 0.7938$.[15] To be 80% confident that the relevant actor would make Choice A is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite *but is otherwise exactly the same as the real world*) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice A. To be 80% confident that the relevant actor would make Choice C is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choice A has already been made) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice C. To be 80% confident that the relevant actor would make Choice E is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choices A and C have already been made) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice E. See Figure 13 in the Appendix for a visual of an example where no Y is more likely than Y.

In a three-step probability tree in which the improbable path leads to Y (Figure 8b), we need to be at least 80% confident that at each decision point, the option that leads to Y will be chosen. To put it another way, to conclude that $P(Y = 1|X = 0) > 0.50$ in a counterfactual world set up as Figure 8b, it must be true that $P_b > 0.7938$, $P_d > 0.7938$, *and* $P_f > 0.7938$.[16] To be 80% confident that the relevant actor would make Choice B is to

[15] Again, technically, there is a number of different combinations of probabilities that could lead to the outcome at the end of the improbable path being greater than 0.50. For example, we could be 79.37% confident that at one of the decision points, the actor chooses the path to no Y, and only 78.75% confident that at another decision point, the actor chooses the path to no Y if we are more than 80.00% confident that at the last decision point the actor chooses the path to no Y. But again, for simplicity's sake, I focus on being at least 80% confident at each decision point.

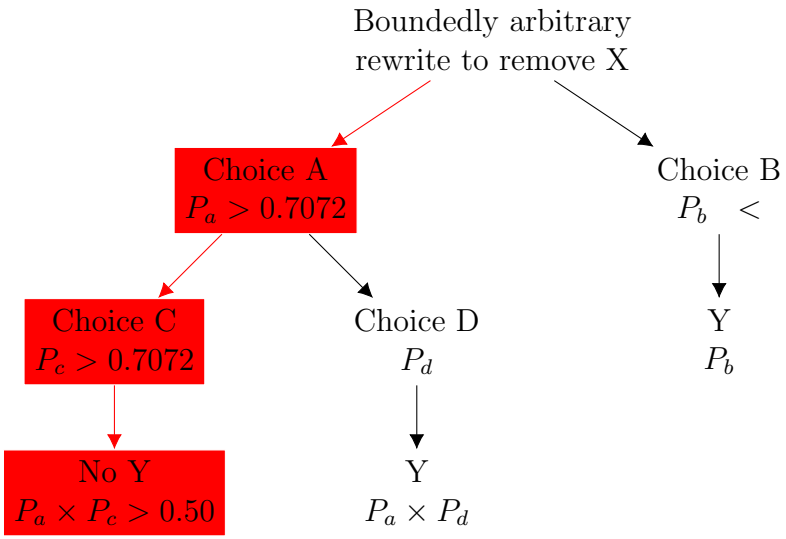
[16] Same as the footnote above.

say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite *but is otherwise exactly the same as the real world*) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice B. To be 80% confident that the relevant actor would make Choice D is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choice B has already been made) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice D. To be 80% confident that the relevant actor would make Choice F is to say that if we could simulate this world (one in which X has been removed by an boundedly arbitrary rewrite and Choices B and D have already been made) 10,000 times over, in at least 7,938 of these worlds, the actor would make Choice F. See Figure 14 in the Appendix for a visual of an example where Y is more likely than no Y.[17]

The math behind the two-step and three-step probability trees dictates *how* confident we must be in our historical evidence at each decision point to conclude whether $P(Y = 0|X = 0)$ or $P(Y = 1|X = 0)$ is greater. Beyond this, however, the math can offer no more guidance. We have no time machine to simulate the world once over, let alone 10,000 times over. What we do have to guide us in how confident we can be, however, is history. The historical evidence can sometimes provide quite strong guidance, especially for measured, considered, debated decisions. Probabilisitic counterfactual analysis should not be used to study decisions that are so trivial that the historical evidence provides no guidance, e.g. what color shirt Khrushchev would wear (Fearon 1996, p. 46). In considering 10,000 counterfactual worlds, Khrushchev could have worn a blue shirt on October 25, 1962 5,000 times and worn a white shirt on October 25, 1962 5,000 times. Historical evidence provides no insight into what we would expect. Neither should probabilisitic counterfactual analysis should be used to study heat-of-the-moment decisions or decisions where accidents, luck,

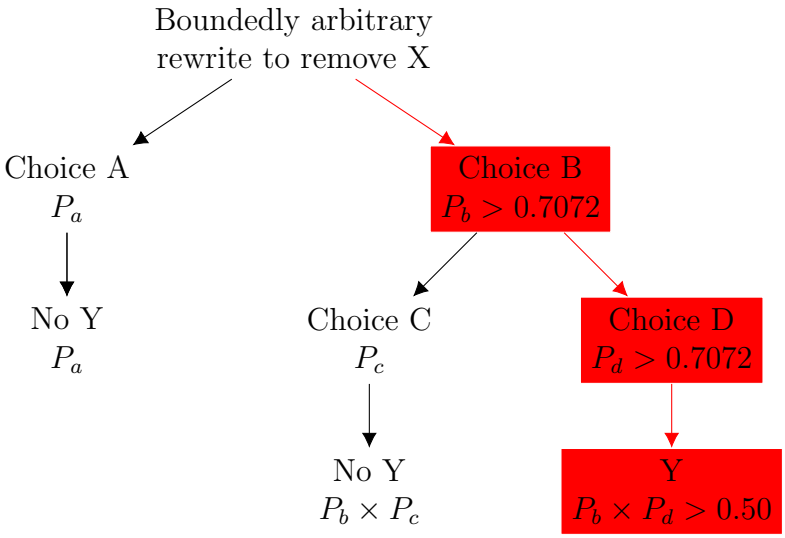
[17] I suggest that the limit for the number of steps should be three steps. There can be more steps, however, the confidence of each step would have to be even higher. For example, in a four-step decision tree, to reach the end of the improbable path, at each decision point, the analyst must be at least 84% confident ($\sqrt[4]{0.50} \approx 0.8409$) that the choice leading to the end of the improbable path would have been chosen. In a five-step decision tree, at least 87% confident ($\sqrt[5]{0.50} \approx 0.8706$), in a six-step decision tree, at least 89% confident ($\sqrt[6]{0.50} \approx 0.8909$), and so forth.

Figure 12: Minimum Level of Confidence in Two-Step Probability Trees



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_a \times P_c = P(Y = 0|X = 0)$$
$$P_b + (P_a \times P_d) = P(Y = 1|X = 0)$$
$$P(Y = 0|X = 0) > P(Y = 1|X = 0)$$

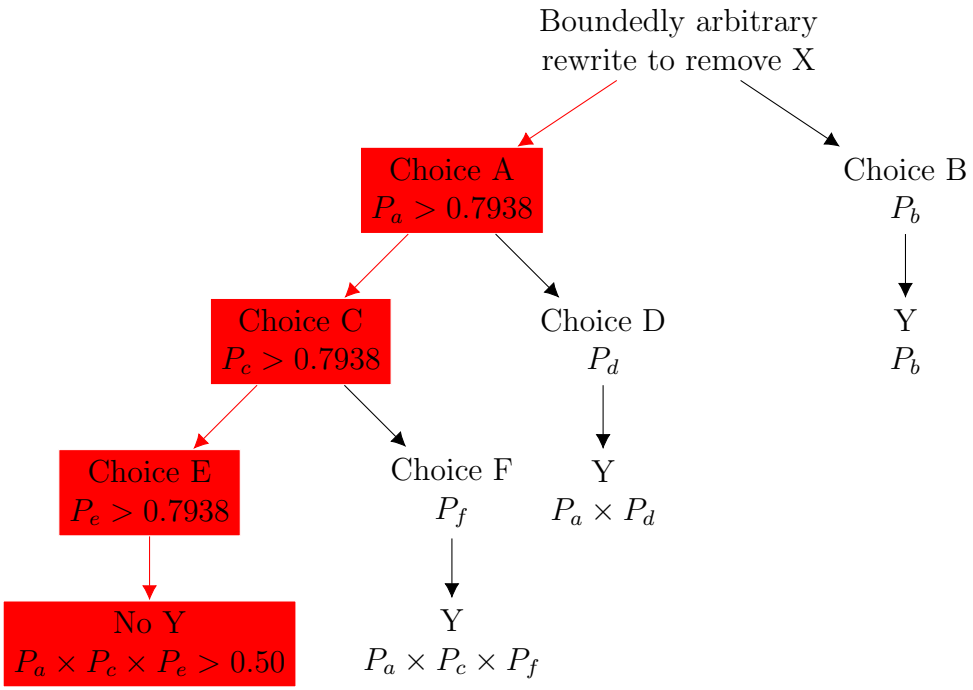
(a) Generic Two-Step Probability Tree 1: Likely Outcome is No Y



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_a + (P_b \times P_c) = P(Y = 0|X = 0)$$
$$P_b \times P_d = P(Y = 1|X = 0)$$
$$P(Y = 0|X = 0) < P(Y = 1|X = 0)$$

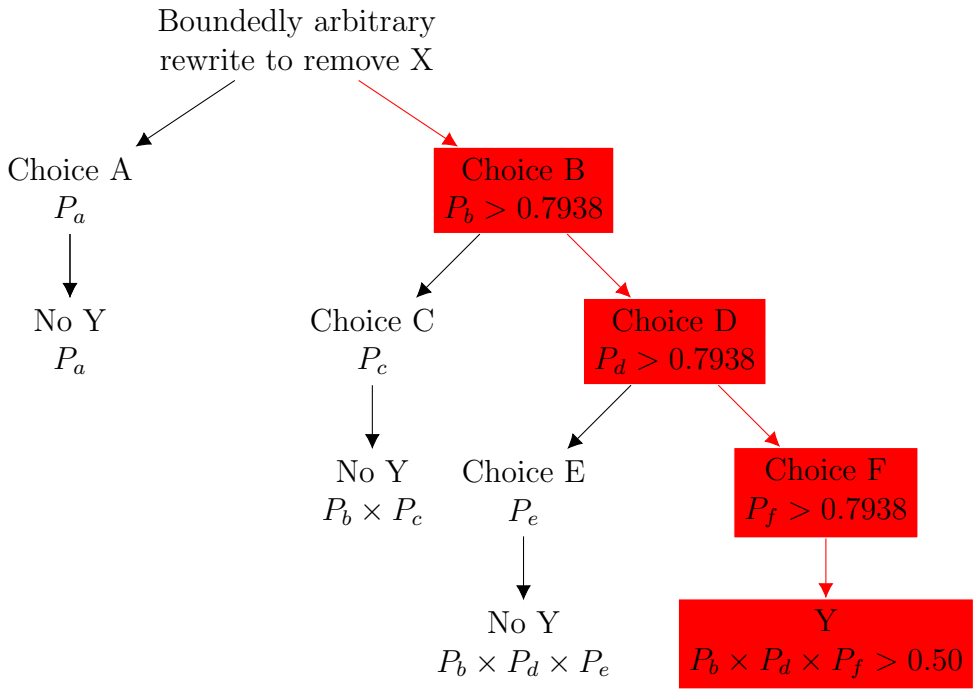
(b) Generic Two-Step Probability Tree 2: Likely Outcome is Y

Figure 13: Minimum Level of Confidence in Three-Step Probability Tree 1: Likely Outcome is no Y



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_e + P_f = 1$$
$$P_a \times P_c \times P_e = P(Y = 0|X = 0)$$
$$P_b + (P_a \times P_d) + (P_a \times P_c \times P_f) = P(Y = 1|X = 0)$$
$$P(Y = 0|X = 0) > P(Y = 1|X = 0)$$

Figure 14: Minimum Level of Confidence in Three-Step Probability Tree 2: Likely Outcome is Y



$$P_a + P_b = 1$$
$$P_c + P_d = 1$$
$$P_e + P_f = 1$$
$$P_a + (P_b \times P_c) + (P_b \times P_d \times P_e) = P(Y = 0|X = 0)$$
$$P_b \times P_d \times P_f = P(Y = 1|X = 0)$$
$$P(Y = 0|X = 0) < P(Y = 1|X = 0)$$

and chance played a large role in which, again, historical evidence may not provide much insight. On October 27, 1962, Vasili Arkhipov, the second-in-command on board the the B-59 nuclear-armed submarine in the vicinity of the US blockade line around Cuba, refused to consent to launching a nuclear-tipped torpedo in response to the USS Beale dropping depth charges. In considering 10,000 counterfactual worlds, it is impossible to estimate in how many of these counterfactual worlds Arkhipov would have been able to calm the captain down and avoid nuclear war, and in how many of these counterfactual worlds the captain might have coerced him to give consent and started a nuclear war.

Probabilisitic counterfactual analysis should be only applied to measured, considered, debated decisions for which there is a paper trail in the real world from which we can extrapolate what would most likely happen in a counterfactual world where X is arbitrarily removed. Depending on the question and the case, sometimes the historical evidence is so clear and so abundant that one would be confident that in considering 10,000 counterfactual worlds, a given decision would be made 9,000 times, 9,500 times, or 9,900 times. However, we do not even need to be that confident. In a two-step probability tree, it is enough to be confident that in 10,000 counterfactual worlds a given decision would be made 7,072 times, and in a three-step probability tree, 7,983 times, at each decision point.

My stopping rule for data collection is when I have collected enough data to determine whether the historical evidence is ambiguous, limited, or clear and abundant enough to meet the minimum confidence level. In other words, I have consulted enough secondary and primary sources to conclude that there is some, perhaps even a lot, of evidence about how an actor handled a certain decision in the real world, but their reasoning is so tied up in the part of the real world that I changed in my boundedly arbitrary rewrite that the real world provides no basis for extrapolating into the counterfactual world ($P = 0.50$). *Or* I have consulted enough secondary and primary sources to conclude that this actor never faced this choice in the real world, so there is no historical evidence from the real world to extrapolate anything about the counterfactual world ($P = 0.50$). *Or* I have consulted enough secondary

and primary sources to be either 71% or 80% confident - depending on the number of steps in the relevant decision tree - about what the actor would have chosen in the counterfactual world ($P \overset{?}{>} 0.7072$, $P \overset{?}{>} 0.7938$).

Omitted Variable Bias

Probabilistic counterfactual analysis, like any other method, cannot completely eliminate the omitted variable bias. The omitted variable bias is the bias that is introduced when the analyst does not consider, i.e. omits, a variable that actually had an important effect on the outcome of interest. In the application of probabilistic counterfactual analysis, the omitted variable bias is when the analyst did not account for an important variable in their counterfactual world because they may be unaware of it or underestimated how important it was. For example, if an analyst is working in an unfamiliar context, they may underestimate the political importance of familial ties in this context, or underestimate the importance of informal leaders in this context. It will be impossible to account for every single variable that is likely to shape actors in the counterfactual world. Analysts will be forced to make decisions about what variables to discuss and consider in the interest of length and clarity of writing. These decisions - about what variables mattered most - are more likely to be sound decisions, and have less omitted variable bias, when the scholar is working in a context in which they are deeply immersed and have a broader understanding of beyond the narrow question they are trying to answer. These same decisions are most likely to be unsound decisions that have more omitted variable bias when the scholar is working in a context in which they are not deeply immersed or are only trying to answer the narrow question in.

My final piece of advice to scholars interested in using probabilistic counterfactual analysis is to apply it to cases in which one is already a country, region, or functional expert in. Good counterfactual analysis, according to Levy and Goertz (2007, p. 398), "includes a significant empirical as well as theoretical component. This requires the analyst to be well immersed in the history of the case and in the secondary literature. The casual interloper cannot possess

this knowledge." If one wants to test the generalizability of an argument to cases outside one's country, region, or functional expertise, I strongly encourage supplementing probabilistic counterfactual analysis with other methods of direct observation.[18] In short, probabilistic counterfactual analysis is a great method for anyone who is already a substantive expert to hone and sharpen how they think about their cases. It is also great as one method in a mixed-methods research design for those who want to understand patterns and relationships in more than just the context one is already a substantive expert in.

The Novelty of Probabilistic Counterfactual Analysis

Existing approaches are *deterministic* because they take a yes-or-no approach to analyze what follows from a counterfactual premise - though they do not say it explicitly, implicitly they are contending that one potential outcome is 100% likely (that the analyst is defending) and the other potential outcome is 0%. My approach is explicitly *probabilistic* - is the likelihood of this potential outcome greater than 50%? If so, by definition, the likelihood of the other potential outcome is less than 50%, and if not, by definition, the likelihood of the other potential outcomes is greater than 50%.

In many existing approaches to counterfactual analysis,[19] the analysis is *the argument.* In my approach to counterfactual analysis, the analysis is *the systematic and disciplined collection and organization of relevant historical evidence to make probabilistic judgments.* In many existing approaches, the rewrite is determined by *historical plausibility.* In my approach, the rewrite is *boundedly arbitrary* - a miraculous rewrite brought about by the most plausible rewriting possible. In many existing approaches, there is *only one consequent for every counterfactual antecedent* under consideration. The validity of the counterfactual analysis rests in the *soundness of the connection from the antecedent to the consequent.* In my

[18] Another option is to collaborate with a coauthor who is interested in the same theoretical question but has a different country, region, or functional expertise.

[19] As covered in my section on existing approaches to counterfactual analysis, there are *many* existing approaches and not all of the distinctions I discuss in this section will apply to *all* existing approaches. However, the differences I highlight are differences between my approach and a typical political science counterfactual analysis in the positivist tradition.

approach, there are *two consequents for every counterfactual antecedent* under consideration. The validity of the counterfactual analysis rests in the *quality of the historical evidence used to calculate the relative probabilities of either potential consequent.* With the exception of Harvey (2012), my approach is novel in how explicitly it treats the existence of two consequents. While Harvey's approach focuses on a *comparing* of the evidence in support of No X-No Y versus No X-Y counterfactuals, my approach focuses on weighing the evidence to generate a *probability distribution* between the two consequents of Y and No Y in a No X counterfactual world.

The benefit of my methodology over existing applications of counterfactual analysis is its transparency. My method forces analysts to be transparent about how confident they are that this actor would have made this choice (50%? Greater than 71%? Greater than 80%?) at each decision point and present their evidence. If others disagree with the analyst's conclusion, then the critics need to bring additional evidence to the table to show that the analyst should be more or less confident than they are about which options would have been chosen at which decision points. It is then possible to reassess which outcome was more likely in light of this new evidence. Sometimes, the new evidence will change which potential outcome is more likely, and sometimes the evidence will not change which potential outcome is more likely. Either way, it allows for a way to structure debates over what would most likely have happened in a transparent way.

Example Application

I use Khong (1996) as an example of how even existing excellent counterfactual studies can be made more transparent through the application of my method. For Khong, the the counterfactual antecedent is "If Britain had confronted Hitler at Munich" and the two consequents he focuses on are "Hitler would have backed down" and "WWII might have been avoided" (p. 95). I reframe Khong's single-case study as one case in a more general question about the role of individual leaders in war onset. Based on Khong's analysis of British

politics before and during Munich, and according to the rules of probabilistic counterfactual analysis I set forth, I construct the following decision tree (see Figure 15).

The improbable path in this decision tree is the path that leads to No WWII. In order to conclude whether No WWII (Y=0) or WWII (Y=1) is the more likely outcome in this counterfactual world, we focus on whether the probability of No WWII is greater or less than 0.50. For the probability of No WWII to be greater than 0.50, because this is a two-step decision tree, we must be at least 71% confident that at each of the decision points, the choice that led to No WWII was the choice that would be made by the counterfactual world leader. In other words, it must be true that $P_a > 0.7072$ *and* $P_c > 0.7072$.

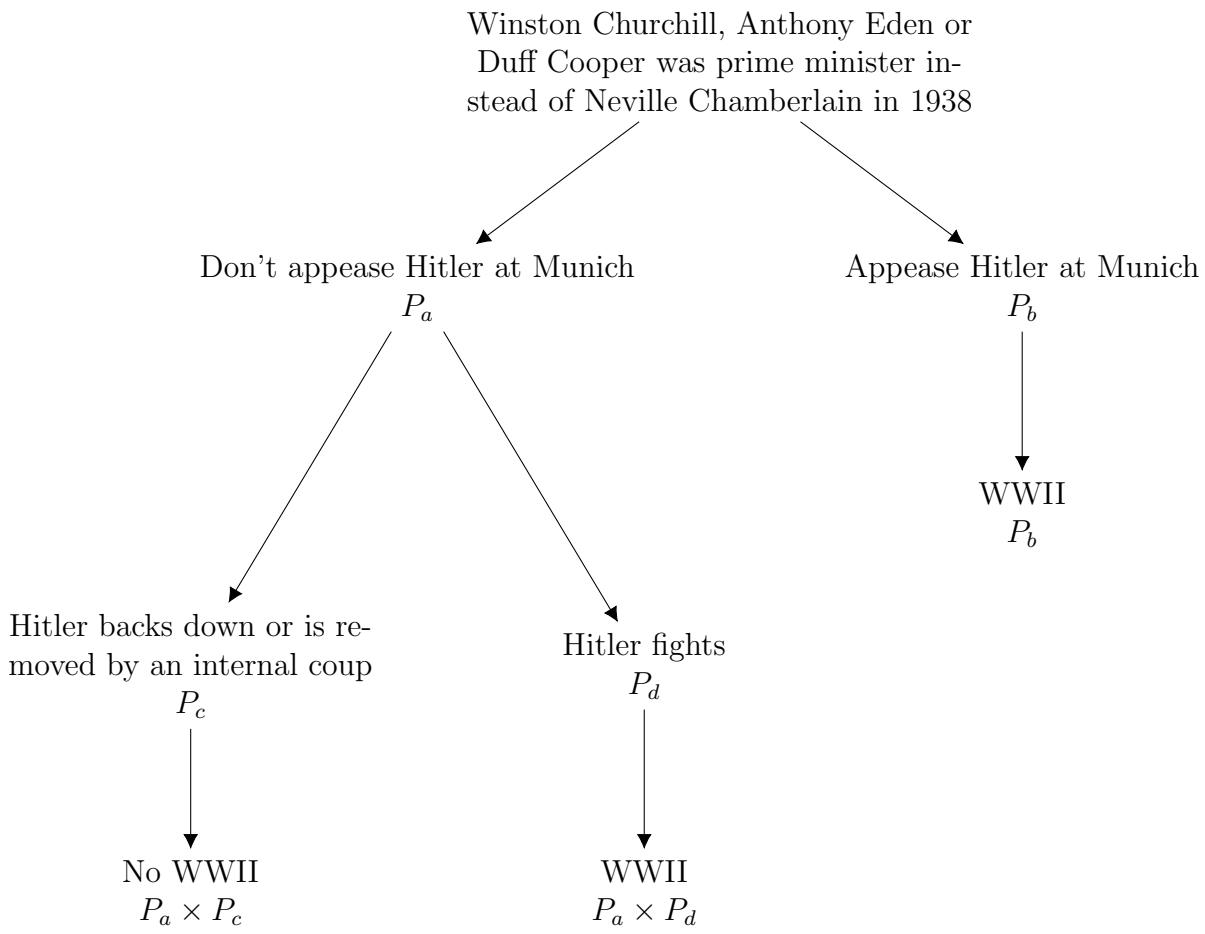
The historical evidence that Khong presents on Winston Churchill (p. 110-111), Anthony Eden (p. 105-107), and Duff Cooper (p. 108-110) makes one fairly confident that in at least 7,102 out of 10,000 counterfactual worlds where everything was the same except one of these three were prime minister instead of Chamberlain, they would choose to *not* appease Hitler at Munich. Therefore, $P_a > 0.7072$. However, the historical evidence about whether Hitler would have backed down or been removed by an internal coup is ambiguous and limited (p. 114-116). Therefore, $P_c = 0.50$.[20] In conclusion, in a counterfactual world where Winston, Eden or Cooper was prime minister instead of Chamberlain in 1938, probabilisitic counterfactual analysis suggests that WWII was still more likely than WWII.

Contribution to the Discipline

This method can be applied to generalizable and narrowly defined questions in political science. I believe it is particularly suitable for studying the role of leaders in cases where there is sufficient data to extrapolate what a different leader would have done and how other actors would have reacted to this counterfactual leader's same or different choices. This method is well-suited to the study of leaders' role in military intervention (Saunders 2011), medium-

[20]Note that for $P_a \times P_c > 0.50$, P_a would have to be infinitely close to 1.

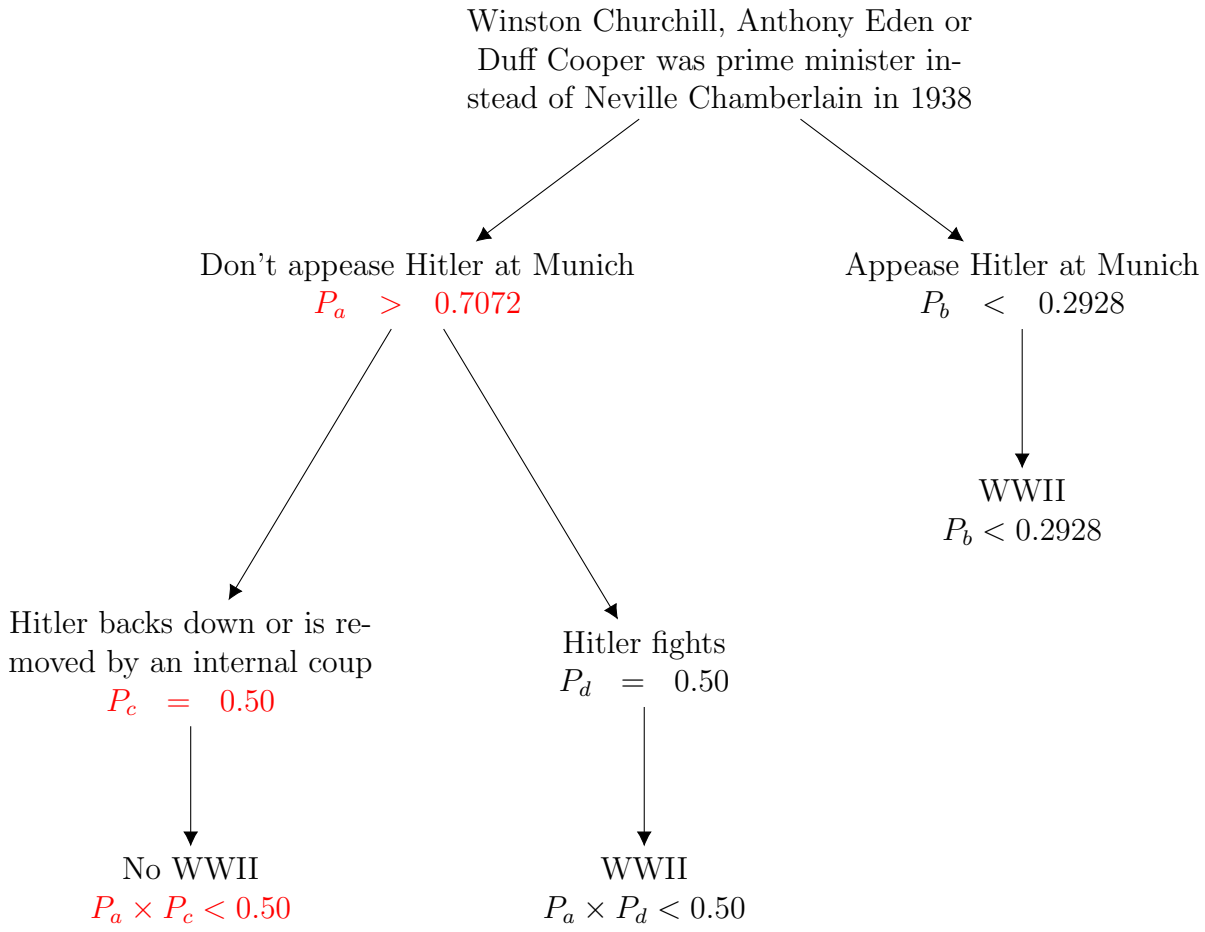
Figure 15: Decision Tree of Khong's (1996) Counterfactual Analysis



$$P_a + P_b = 1, P_c + P_d = 1$$
$$(P_a \times P_c) + (P_a \times P_d) + P_b = 1$$
$$P_a \times P_c = P(Y = 0|X = 0)$$
$$(P_a \times P_d) + P_b = P(Y = 1|X = 0)$$

$X = 1$: Chamberlain was British prime minister in 1938 (Real world premise)
$X = 0$: Churchill, Eden, or Cooper was British prime minister in 1938 (Counterfactual world premise)
$Y = 1$: WWII (Real world outcome, potential counterfactual world outcome)
$Y = 0$: *No* WWII (Potential counterfactual world outcome)

Figure 16: Probability Tree of Khong's (1996) Counterfactual Analysis



Historical evidence suggests that $P_a > 0.7072$ and $P_c = 0.50$
Therefore, barring $P_a = 1$, $P_a \times P_c < 0.50$

$$P(Y = 0|X = 0) < 0.50$$
$$P(Y = 1|X = 0) > 0.50$$
$$P(Y = 0|X = 0) < P(Y = 1|X = 0)$$

Conclusion: WWII would more likely have started than not even if a Churchill, Eden, or Cooper was prime minister instead of Chamberlain.

term causes of war (Kennedy 2011),[21] weapons development and nuclear proliferation, and war termination, and for long, drawn out negotiations in peacetime (Walker 1977), where the process of negotiation itself generates a lot of data about the preferences of plausible leaders. In particular, this method could be applied well to the question of leaders' role in trade agreements, conventional and nuclear arms limitation agreements, and the formation of international regimes. The method could also be used to understand leaders' role in alliance formation, states honoring or not honoring alliance commitments, and in alliance renewal and alliance termination.

[21]This method is not suited to studying leaders' role in long-term causes of war, because there are too many steps between the counterfactual antecedent and hypothesized consequent. This method is not suited to to study leaders' role in the short-term causes of war or decisions in immediate crises, because the "fog of war" would make it difficult to predict how individual leaders would behave had they been under stresses that are incomprehensible to those not under the same stresses.

References

Brady, Henry E. 2008. *The Oxford Handbook of Political Methodology*. Oxford, UK: Oxford University Press chapter Causation and Explanation in Social Science, pp. 217–270.

Capoccia, Giovanni. 2016. *Critical junctures*. Oxford University Press Oxford.

Capoccia, Giovanni and R Daniel Kelemen. 2007. "The study of critical junctures: Theory, narrative, and counterfactuals in historical institutionalism." *World politics* 59(3):341–369.

Collier, David and Gerardo L Munck. 2017. "Building blocks and methodological challenges: A framework for studying critical junctures." *Qualitative and Multi-Method Research* 15:2–9.

Collier, Ruth Berins and David Collier. 1991. *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America*. Princeton, NJ: Princeton University Press.

Fearon, James. 1996. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton, NJ: Princeton University Press chapter Causes and Counterfactuals in Social Science: Exploring an Analogy between Cellular Automata and Historical Processes, pp. 39–67.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43(2):169–195.

George, Alexander L and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Goodman, Nelson. 1947. "The Problem of Counterfactual Conditionals." *The Journal of Philosophy* 44(5):113–128.

Goodman, Nelson. 1983. *Fact, fiction, and forecast*. Harvard University Press.

Greenstein, Fred I. 1992. "Can Personality and Politics Be Studied Systematically?" *Political Psychology* 13(1):105–128.

Harvey, Frank P. 2012. *Explaining the Iraq War: Counterfactual Theory, Logic, and Evidence*. New York: Columbia University Press.

Jervis, Robert. 2013. "Do leaders matter and how would we know?" *Security Studies* 22(2):153–179.

Kennedy, Andrew. 2011. *The International Ambitions of Mao and Nehru: National Efficacy Beliefs and the Making of Foreign Policy*. Cambridge, UK: Cambridge University Press.

Khong, Yuen Foong. 1996. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton, NJ: Princeton University Press chapter Confronting Hitler and Its Consequences, pp. 95–118.

Lebow, Richard Ned. 2010. *Forbidden Fruit: Counterfactuals and International Relations*. Princeton, NJ: Princeton University Press.

Levy, Jack. 2008. *Oxford Handbook of Political Methodology*. New York: Oxford University Press chapter Counterfactuals and Case Studies, pp. 627–44.

Levy, Jack and Gary Goertz. 2007. *Explaining war and peace: Case studies and necessary condition counterfactuals*. New York: Routledge.

Lewis, David. 1973. "Causation." *Journal of Philosophy* 70:556–67.

Saunders, Elizabeth N. 2011. *Leaders at War*. Ithaca, NY: Cornell University Press.

Tetlock, Philip E and Aaron Belkin. 1996. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton, NJ: Princeton University Press chapter Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives, pp. 3–38.

Thelen, Kathleen. 1999. "Historical institutionalism in comparative politics." *Annual review of political science* 2(1):369–404.

Walker, Stephen G. 1977. "The interface between beliefs and behavior: Henry Kissinger's operational code and the Vietnam War." *Journal of Conflict Resolution* 21(1):129–168.

Weber, Max. 1949. *The Methodology of the Social Sciences*. New York: Free Press.