

UNITED WISDOM

Counterfactual explanation: what would change under a different condition?

World and Consciousness

Andrey Zarinsh & Maria Evilkut

W-13-N02 · EN

@andrey.zarinsh @ilkutki

Counterfactual explanation: what would change under a different condition?

Epigraph

> *“[…] if the first object had not been, the second never had existed […]”*

David Hume, An Enquiry Concerning Human Understanding, E 7.29, SBN 76–77. A checked conditional clause with marked omissions. It belongs to the first of two formulations of cause, alongside conjunction and customary expectation. It is not an independent modern method for identifying outcomes. Context.

What this article is about

After an event, it is easy to say, “If one condition had been different, everything would have happened differently.” The sentence can help an investigation or remain a convincing story. We know the outcome that occurred; we do not have a second outcome under those same circumstances before us. The question concerns the grounds connecting a changed condition with a changed consequence, rather than the attractiveness of an alternative history.

This article compares specified passages from Hume and Mill with a bounded account of Judea Pearl’s structural models. Hypothetical digital-indicator and meeting examples examine what stays fixed, why observation differs from intervention, and when a conclusion remains conditional. Neither example is an experiment on people or an account of their blame. The final practice suggests recording one counterfactual question: the precise change, presumed constants, grounds for the conclusion, and unknowns. Such a record promises neither to solve every cause nor to rewrite what already happened.

I. A question about another outcome

Imagine a small meeting beginning later than planned. A participant says, “If the invitation had been sent yesterday, there would have been no delay.” This proposes an explanation by changing one condition: the invitation’s timing. It does not yet show why that timing would have ensured the desired result. The invitees might still have been busy; the room might have become available only later. These are hypothetical alternatives, offered to show that a counterfactual sentence needs specified retained conditions, rather than merely a different event in its first clause.

@andrey.zarinsh @ilkutki

Defining the outcome is a useful start. “Everything would have been different” is too broad: an earlier invitation already changes the history, but our concern may be only an on-time start. A more precise question is, “With an earlier invitation, would this meeting have begun at the appointed hour?” We separate a difference in the overall story from a difference in the selected result. Its grounds can then be discussed. We need the invitation’s relationship to participants’ availability and the location, rather than a list of everything that might have differed in their lives.

A counterfactual concerns an option not realised in this case. It differs from a report comparing two meetings that actually occurred. Such a comparison may provide some grounds if its differences and limits are explicit. We cannot substitute a similar observed case for the missing alternative and declare its circumstances identical. Nor is every question without a second observation meaningless. Between direct observation and arbitrary fantasy lies conditional reasoning based on specified relationships. Its conditional status must remain visible, along with the possibility of examining its assumptions rather than treating them as established facts by default.

The question may serve a practical aim: decide what to change when organising the next meeting. Explaining the past and planning the future are nevertheless different. A new day brings other conditions, which the proposed relationship may not cover. It helps to finish the first record with a particular event, criterion of delay, and change in invitation timing. A future action can then be considered separately. This distinction preserves the relevance of experience without turning an interesting alternative account into a ready-made guarantee that the next attempt will succeed under different circumstances.

II. Hume: without the first, without the second

In E 7.29 of the Enquiry, Hume offers two formulations of cause, relating them to observed conjunction and a transition of thought. The first includes a phrase stating that without the first object, the second would not have existed. This is a historical philosophical passage, not an algorithm for calculating individual outcomes. We use it to examine the question “what would happen without this condition,” while preserving Hume’s discussion of customary expectation. It supplies no data about our hypothetical meeting and cannot replace an investigation of its particular circumstances.

Consider a hypothetical digital indicator. Its rule is specified in advance: it displays one only when both inputs A and B equal one; otherwise it displays zero. In the case under discussion, both inputs and the output equal one. Changing A to zero while retaining B and the rule changes the output to zero. This is a deduction from an explicitly defined toy model. It does not establish how a real device works. Its benefit is that the reader can see the grounds of the answer and the precise change being considered.

Removing A changes the result here, but A is not the only condition. If B were zero, setting A to one would still not produce one. “Without A there is no result” therefore does not imply that A alone suffices. Necessity in this comparison differs from the sufficiency of a single condition. The meeting is more complicated: an earlier invitation may help without ensuring that participants or a room are available. The indicator analogy clarifies the structure of the question. It does not prove that human action follows the simple rule selected for a model we invented.

The direction of explanation also matters. A known result and an assumed rule can be used to investigate compatible inputs, but the result does not always identify a unique history. Even our indicator’s zero permits several states: A missing, B missing, or both. The possibilities can be enumerated for this rule. A real event requires separate information about the actual conditions. A counterfactual conclusion cannot therefore rest solely on the impression that an absent result must mean the absence of whichever condition currently seems most convenient to name as its cause.

III. Mill: comparison must preserve the relevant differences

In Book III, Chapter 8, §2 of A System of Logic, Mill formulates the Method of Difference. The cases must share relevant circumstances except the selected difference. His canon allows not only a cause but an indispensable part of a cause. The selected §3 context discusses strict similarity’s difficulty and the possibility that the procedure itself introduces another change. Historical dangerous examples are not adopted as instructions here. The comparison’s conditions matter: a changed result does not establish that the intended difference was the only one present.

Suppose the digital indicator is checked twice. Initially A and B equal one; then A equals zero while B remains one. If the rule really stays unchanged, this comparison addresses our question. If the second trial also changes the rule or B, its result cannot be attributed solely to A. “We changed one input” requires grounds for saying that the rest was retained. A model states that as a condition; observation of a real device makes it a separate part of the work to be checked, rather than something that follows from the investigator’s intention alone.

Meeting days with early and late invitations could be compared, but they differ in more than invitation timing. Schedules, participants, room availability, and waiting time may also change. A connection between invitation timing and the start still needs an account. Early invitations might have been used for meetings with rooms secured in advance. The selection condition would then relate to the outcome. This hypothetical alternative is not a claim about what usually happens. It shows why matching one conspicuous detail does not guarantee that other relevant circumstances have remained the same between the cases being compared.

The Method of Difference is useful as a requirement for an inference's grounds, rather than a name that automatically turns a comparison into proof. We can record shared conditions, differences, and unknowns. An unknown relevant difference should remain a limit on the conclusion. Sometimes a model check or ordinary information gathering is available; sometimes an exact comparison is not. Lacking a final answer is then preferable to arbitrary certainty. We preserve the explanatory question without declaring hidden differences eliminated merely because a convenient narrative has been organised around one cause that seems especially prominent to its teller.

IV. A conditional model makes its answer inspectable

In his 2009 overview, Pearl describes structural counterfactual reasoning, §3.4: replace a selected equation with an intervention while retaining other mechanisms and the background context. The answer then has explicit assumptions. This is an author's mathematical approach, not a meeting experiment. Observation alone does not identify every required counterfactual relationship. We reproduce no proofs and infer no philosophical consensus. We use only the bounded structure of the question: what changes, and what is presumed invariant.

For our indicator, a table states that A=1 and B=1 produce 1, while either zero input produces 0. Within that rule, the answer about changing A is transparent. Every table row can be checked, revealing why the assumption about B matters. If someone instead specifies "one whenever either input is one," the answer changes. Identical question wording therefore does not ensure an identical result. It must be accompanied by the relationship retained in this particular argument, rather than leaving the reader to imagine whichever relationship makes the preferred answer easiest to obtain.

Conditional reasoning is different from discovering how the world works. We invented a rule and deduced a consequence. An actual indicator would require grounds for regarding that rule as appropriate. A different mode, another input, or an observation error might be found. Internal table consistency does not establish a model's fit to its subject. The table nevertheless clarifies that question: instead of general confidence, we have a specific claim that can be compared with information about the subject. The deduction remains useful while retaining the condition on which its application depends, rather than disguising it as an observed result.

Our meeting description generally supplies no exact table of this kind. We can suppose that an earlier invitation would have secured participants' responses, but that is a separate hypothesis. If participants say that the appointed time would still have been occupied, we have grounds to reconsider it. Without a response, the question stays open. Missing data cannot be filled with the toy indicator's rules. The example distinguishes deduction from a specified relationship and a claim that the relationship operates in a human situation with its own context, participants, and decisions that have not been replaced by our assumptions.

V. Observing a condition differs from changing it

Consider a separate example. The recorded states always pair A=0 with B=0 and A=1 with B=1. Their outputs are zero and one respectively. These records fit both “the output repeats A” and “the output repeats B.” Both rules answer identically for the observed states. The records alone therefore cannot discriminate between them merely because the output matches A. Additional grounds are needed, and this description supplies none. The example’s point concerns precisely what is absent from this set of observations, rather than a claim that observations can never contribute to causal knowledge.

Now imagine changing A from one to zero while retaining B=1. The first rule gives zero, the second one. Identical observed records permit different answers to the new question. This deduction uses two explicitly proposed toy rules; no experiment has been performed. It identifies a particular insufficiency: the records omit a state that could distinguish the relationships. Greater confidence in one account adds no missing information. We must justify the rule or find an acceptable way to clarify the question about the subject. Neither task is completed by naming the preferred rule more emphatically.

Invitation timing can likewise be observed alongside other conditions. Early invitations and on-time starts may coincide without their relationship being established. Both might depend on organisation completed in advance. Would changing only the sending time preserve the result? That is a new question. Observing people who received early invitations differs from changing the invitation time for this meeting. The hypothetical example proposes no intervention without consent. Its purpose is to name a missing relationship rather than authorise anyone to control another person’s behaviour or presume the response that a participant should give.

Some relevant information may be obtainable through an ordinary question or an organisational record; other questions may remain unsettled. This boundary is not a defeat of reasoning. We have established what is insufficient: one coincidence between A and the output within a limited set of cases. That is a substantive conclusion about these grounds. It can prevent a premature alternative history and guide a further check. It does not prove that A has no influence whatsoever, or that a counterfactual account of this subject could never receive justification through additional information.

VI. A group answer is not an answer about this case

A counterfactual can concern one event or a group. “Would this meeting have begun on time?” differs from “Do meetings generally start earlier with this invitation procedure?” The second question needs a specified group, criterion of starting, and comparison method. Even a useful group answer need not establish this meeting’s alternative outcome. An unavailable room does not become accessible through an average improvement in invitations. The distinction is already present in the subject of the question, before selecting a sophisticated formula or assigning a numerical estimate to an insufficiently specified comparison.

In our two-required-input model, changing A from zero to one has different effects depending on B. With B=1, the output changes from zero to one. With B=0, it remains zero. We could define a group with different B values and calculate a chosen fraction of changes. That fraction would not tell us the B value of an unexamined individual case. An individual inference needs the relevant condition or an explicitly conditional answer. Any numbers in this example would concern only the stated collection of states, rather than a discovered distribution of real-world meetings.

The opposite mistake turns one successful case into a rule for everyone. Changing A worked in one state with B retained. That does not show identical effects across every input, rule, and time. In III.10, §§1–2, Mill considers several routes to an effect and limits of the Method of Agreement. His historical context is retained; medical examples are not adopted as guidance, and the selected passage is not a complete modern textbook. It helps distinguish an inference concerning one comparison from a universal claim about every possible case.

A speaker can name the scale directly: “The records suggest this procedure helped some meetings; for this meeting, room availability remains unknown.” This is not a weaker substitute for confidence, but a differently bounded result. It retains the identified relationship and the missing individual information. Repeatability and an individual case can be investigated separately. We cannot move between them without acknowledging a changed question. A group account does not replace one event’s history, while an individual history cannot become a rule for an entire class merely by being told in a persuasive way.

VII. Retaining a mechanism does not freeze its consequences

“Everything else stays the same” needs clarification. Does it retain starting conditions, the relationship’s rule, or every later outcome? These versions can pose different questions. Consider another toy model: B copies A, and Y copies B. Initially all three equal one. Changing A to zero while preserving both copying rules makes B and Y zero. Changed consequences do not violate retained rules; they follow from the retained relationship. The example supplies its rules explicitly, so its inference concerns those stated rules rather than an assumed mechanism of any real meeting or device.

If B is additionally forced to remain one, Y remains one too. That is a different condition: we have changed A and the way B is produced. Both deductions can be correct for their respective toy descriptions. The error lies in quietly replacing one question with another. “With a different A” resembles “with a different A but B forcibly retained,” yet they prescribe different changes. A real task requires grounds for the conditions we retain and for consequences that may change along with the selected factor. We cannot obtain those grounds merely by describing an alternative as otherwise identical.

An earlier invitation might change response timing, preparation, or a location choice. Our example does not establish that these consequences would occur; the information is absent. But we cannot use a proposed chain as an explanation and then forbid its consequences by saying “everything else is unchanged.” The question must be selected: change one organisational action together with its possible consequences, or change it under additionally fixed conditions. Each version has its own assumptions and inference limits. Stating the version lets other participants examine the question actually being asked, rather than supply different versions of it without noticing.

Clarification also prevents an excessively flexible alternative history. Changing the invitation, participants’ schedules, room access, and their decisions involves several changes. The story may describe an intelligible scenario without isolating one condition’s influence. It can be named honestly as a combined proposal and examined separately. It should not be presented as an answer to the earlier narrow question. Preserving the comparison’s boundaries gives counterfactual reasoning substance: the reader can see the selected option, the grounds of the answer, and the portion that depends on assumptions added to obtain the preferred scenario.

VIII. A causal question does not settle blame

“If you had acted differently” can combine an explanation with an accusation. A counterfactual relationship and a moral judgment need different grounds. Even an established changed outcome does not by itself establish available information, authority, obligations, or other contributing conditions. This implies neither that responsibility is always absent nor that it is equally divided. We separate tasks: describe a changed condition’s relationship to a consequence, establish the actual action, and, where needed, discuss evaluation on grounds relevant to that evaluation. None of these tasks is completed by quietly merging it with the others.

In III.5, §3, Mill relates the selection of one named condition to a discussion’s purpose; his casting-vote example specifically notes an emphasis on responsibility. This supplies no modern legal formula. It helps reveal that selecting a “main cause” can depend on the question for which a condition was singled out. Our indicator requires A and B together. Symmetric inputs do not establish symmetric human responsibility. That judgment requires separate information and norms, which a toy rule does not contain and cannot manufacture through an analogy with the people being discussed.

The hypothetical delay may have an organisational cause and several acceptable remedies. Its circumstances must be established before linking it to someone’s personal quality. Counterfactual narratives are particularly vulnerable to hindsight: after the result, something unknown beforehand can seem obvious. We can ask instead which option was available then and what information existed. An answer may remain incomplete. That incompleteness justifies neither arbitrary accusation nor automatic cancellation of responsibility because many circumstances were involved. It marks a boundary that a fair discussion should retain rather than use selectively for whichever conclusion someone already prefers.

A precise statement can help: “We are examining whether invitation timing would have affected the start; this inference does not yet evaluate a person.” The discussion need not smuggle in a moral verdict. Organisational changes can be considered separately, participants heard, and unknowns named. In a substantial conflict, the proposed logical record cannot replace the human discussion required. Its purpose is narrower: distinguish an explanatory relationship, knowledge of an action, and an evaluation. Neither a physical analogy nor a philosopher’s quotation authorises control of another person’s answer or establishes that the person has been fully explained.

IX. A conditional explanation can end with an honest boundary

Counterfactual reasoning need not end with one confident answer. Its result can be conditional: “Given this rule and retained B, changing A changes the output.” It can identify uncertainty: “The records fit two rules with different alternative answers.” Or it can produce a further question about a relevant condition. All three are substantive. Unlike “everything could have been different,” they specify a subject, relationship, and evidential boundary. The reader can see where an answer has been obtained and where an account remains a proposed version requiring more information before it can describe the actual event.

A small organisational choice for the next meeting need not claim a final explanation of the previous delay. Establishing room availability and asking about a suitable time may be reasonable for the new task. They need not be justified by asserting that the old delay was fully explained. A reversible action and an established counterfactual inference are different results. Keeping that boundary allows organisational improvement without pretending to know the past’s single true alternative history. The action’s suitability nevertheless requires participants’ consent and the concrete constraints of the proposal being considered now, rather than relying on its general appeal.

We can also stop when the required comparison is unavailable or unnecessary for the current decision. Missing a final alternative history does not oblige anyone to enumerate possibilities indefinitely. Missing information can be named, an established relationship retained, and another question considered. This is the article’s proposal about proportionate checking, not a measured psychological result. None of the sources establishes this exercise’s efficacy for regret, relationships, or learning. A reader who uses the record chooses a personally suitable way to clarify one question, with the freedom to adapt it or to find it unnecessary for the task.

A counterfactual becomes more inspectable when four parts are visible: the event, precise change, retained conditions, and grounds connecting change with consequence. A fifth part—the open boundary—prevents a conditional answer becoming a universal law. These parts do not guarantee truth by themselves. They make a claim available for discussion and revision. We can deduce within a defined model, cautiously use information about an actual event, and leave the unknown unsettled. Interest in what might have changed can remain without requiring us to declare every intelligible alternative a proven account of how history would have unfolded.

What this means in practice

1. Select a small hypothetical or ordinary organisational example that does not require evaluating someone: message timing, a record-checking order, or room availability. Write the actual result and one precise feature you want to explain. Replace “everything would have been different” with the particular change of interest—for example, the start time rather than the meeting’s entire history. A major personal conflict or health issue is unnecessary for examining the structure of a counterfactual question and distinguishing its elements from the conclusions someone might prefer to reach.

2. Specify the change: which condition takes another value, and when? List two or three relevant conditions presumed retained. Separate starting circumstances from later consequences. Forcibly retaining an outcome that might depend on the changed condition is an additional assumption; name it explicitly. A combined change can also be discussed, but should not be labelled as changing only one factor. If important information is missing, retain the gap rather than filling it with an assumed answer that happens to make the alternative scenario more convenient or convincing.

3. State the relationship’s grounds. Are they a toy model’s stipulated rule, checked subject information, a case comparison, or a guess? For a comparison, identify a relevant difference that may remain unknown. For a model, separately ask why it fits the actual subject if you intend to apply it beyond the hypothetical example. Assign no confidence percentage without grounds. An honest record can say, “These data do not yet distinguish the two accounts,” or “The answer follows only with this retained condition,” without pretending that a stronger conclusion has been obtained.

4. Write a bounded conclusion and the next acceptable question. Preserve “under this rule” and “in these circumstances” when the answer is conditional. If a participant’s response is needed, allow refusal or clarification. New information can revise the corresponding conclusion without rewriting the whole history for convenience. The record may end at an identified boundary. It offers a way to clarify reasoning without a demonstrated psychological effect; it establishes neither another person’s blame, freedom, nor inner workings. You can stop when the question is sufficiently clear for the task that prompted it.

Key quotation

> *“[…] this method compares an instance of its occurrence with an instance of its non-occurrence […]”*

John Stuart Mill, A System of Logic, III.8, §2. A checked subject-and-predicate clause with marked omissions. “This method” means the Method of Difference; “its” refers to the

phenomenon. The neighbouring canon requires shared circumstances except the selected difference and allows an indispensable part of the cause. Context.

Connections between traditions

Comparison	Working connection	Boundary
Hume, E 7.29, without the first / Mill, III.8, §2, canon	A question about removing a condition needs specified comparison conditions.	A philosophical formulation is not an executed Method of Difference; an indispensable part is not the sole cause.
Hume, E 7.28, repetition and expectation / Mill, III.8, selected §3	A familiar conjunction does not establish the absence of other changes in a check.	Expectation's origin and requirements for comparison concern different tasks, not one theory.
Hume, E 7.29, counterfactual phrase / Pearl, §3.4, definition	An alternative answer needs specified grounds.	A historical formulation and a structural model have different assumptions; Pearl does not calculate this meeting.
Mill, III.8, §2 / Pearl, §3.2.1, intervention	Specify the change and what is retained.	Case comparability and equation replacement differ; model fit needs separate grounds.
Hume, E 7.29, first and second / Mill, III.10, §§1–2	Examine alternative routes before calling one condition universally necessary.	A counterfactual definition and plural causes are not one rule; an individual case differs from a whole class.
Hume, E 7.30, customary inference / Mill, III.5, §3, selecting a named condition	A convincing cause label still needs grounds and a stated explanatory purpose.	The origin of an idea and selection for explanation or responsibility are different questions; moral evaluation remains separate.

Six original comparisons, without a shared-doctrine, historical-borrowing or tested-practice claim.

Primary sources

• David Hume, An Enquiry Concerning Human Understanding, E 7.28–7.30, original English, Hume Texts Online, E/SBN. These paragraphs read completely; the whole VII is not claimed. The page's print edition was not established here.

• John Stuart Mill, A System of Logic, eighth edition, Harper & Brothers, 1882. Complete III.5, §3; III.8, §2; III.10, §§1–2 read, with the specified bounded III.8, §3 context. Historical examples are not contemporary practical guidance.

• Judea Pearl, Causal inference in statistics: An overview, Statistics Surveys 3, 2009, 96–146, DOI 10.1214/09-SS057. Specified intervention definitions and complete §3.4 read; remaining paper and proofs not reproduced. A primary methodological text describing the author's framework.

Tags

counterfactual explanation, cause, condition, consequence, change, comparison, intervention, observation, model, rule, background context, assumption, data, uncertainty, indicator, necessary condition, sufficient condition, plural causes, individual case, group, repeatability, Hume, Mill, Pearl, Method of Difference, responsibility, blame, history, planning, inference limits

About the authors

Andrey Zarinsh is the author and founder of United Wisdom. He explores inner transformation, the unity of traditions, intercultural dialogue, and contemporary spiritual discipline. Instagram: @andrey.zarinsh.

Maria Evilkut is a co-author of United Wisdom. Alongside her research, she seeks to test these ideas in practice and observe how principles of unity and mutual understanding work in real life between people and cultures. Instagram: @ilkutki.