CAUSAL MAP GARDEN

Task 2 – Causal coding – minimalist style

Contents

Task 2 – Introduction

Chapter contents.

9 Oct 2025

> Standing on the shoulders of giants

In this chapter we present some of key general principles about how to do causal mapping which we at Causal Map Ltd (and, most of the time, at BathSDR) have adopted.

This is a very restricted yet powerful **minimalist** approach which we have also called "barefoot" or "naïve" coding.

In the next chapter Tasks 2 & 3 — Extensions — Introduction we look at specific conventions to make causal coding simple and powerful.

PAGES IN THIS CHAPTER

- A minimalist approach to coding does not code absences
- Our approach is minimalist – we do not code the strength of a link
- In a causal mapping dataset there is no need for a special table of factors
- Factor labels – a creative challenge
- Factor label tags – coding factor metadata within its label
- Factor labels – semi-quantitative formulations can help
- Causal mapping looks for linearity first
- Factor labels – do not over-generalise
- Coding with and using link metadata
- Link metadata – Sentiment
- Link metadata – Time reference
- Link metadata – quality of evidence
- Research on the ability of LLMs to detect causal claims

🌻 Our approach is minimalist – we code only bare causation

9 Oct 2025

Why we stick to bare causation in causal mapping.

Our rule of thumb: record only that "C causes D." No coding of necessity, non-linearity, moderators, or strength. Just who said what causes what.

The short case

- **It avoids false precision.** Labelling links as "necessary," "moderator," "non-linear," or assigning strengths suggests evidence we rarely have. We prefer to show what was claimed and how often, then let readers judge. Maps are primarily **epistemic**—repositories of evidence about people's beliefs—not truth machines.
- **It scales and compares.** Bare links plus rich factor labels let us aggregate, filter, and compare across sources, groups, and contexts without fighting about semantics of special symbols. Our tools then summarise with counts (citations, sources) and simple derived measures (like "outcomeness"), instead of speculative link attributes.

What we record

- **Factors (boxes):** short propositions that do the heavy lifting (e.g., "Not enough money," "Won't take a holiday this year").
- **Links (arrows):** undifferentiated causal influence claims between factors. A link means "P said C influences D." That's it.

What we deliberately don't code on links

- Necessity/sufficiency
- Non-linear forms or feedback classifications
- Moderator/mediator/inhibitor role
- Polarity or strength

Why? Because (a) respondents seldom state these explicitly; (b) analysts rarely agree on them from text alone; and (c) they reduce inter-coder reliability and slow projects down without very much which we can dependably aggregate.

Our analyses are still useful

Coding bare links doesn't make maps "impoverished": Causal mapping produces models you can query to answer questions

Bottom line

Most of the time, we code only: "C causes D (as claimed by P)." That minimal, transparent unit is reliable, scalable, and faithful to the data people actually provide. Everything richer belongs in **analysis and interpretation**, not in speculative link types baked into the coding.

🌻 Our approach clearly distinguishes evidence from facts and does not automatically warrant causal inferences

> From Better Evaluation.

Causal mapping distinguishes carefully between evidence for a causal link and the causal link itself. It does not provide any specific way to make causal inferences from one to the other. Causal mapping can help the evaluator to identify, code, simplify and synthesise the evidence for causal connections, but the evaluative step to make a judgement about whether one thing in fact causally influences another is left to the evaluator.

But, Causal Mappers are like Janus

> From (Powell et al. 2024)

..., like Janus, the causal mapper looks in two directions at once: sometimes interpreting maps as perceptions of causation but also often wanting to make the leap to inferences about actual causation. As Laukkanen and Wang (2016: 3) point out, while conceptually poles apart, in practice, the two functions can be hard to distinguish, particularly without sufficient explanation about source information and how this has been analysed. Historically, many causal mappers have been happy with this dual focus and moving from one to the other.

As evaluators, we try to be more rigorous about this distinction. We see the job of the causal mapper as being primarily to collect and accurately visualise evidence from different sources, often leaving it to others (or to themselves wearing a different hat) to draw conclusions about what doing so reveals about the real world. This second interpretative step goes beyond causal mapping per se (Copestake, 2021; Copestake et al., 2019a; Powell et al., 2023).

Relevant page: The elephant in the room — causal inference ▸

References

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

🌻 Our approach is minimalist – factors are not variables

23 Sep 2025

Many or most causal mapping approaches, including Causal Loop Diagrams, also code the perceived strength of a causal link. This means that the factors become variables which can take values between, say, low and high or positive and negative, and we can make a much broader range of inferences using some form of numerical modelling. This can be seen as the extreme reproducible end of our spectrum and borders on quantitative approaches.

However we do not go so far: our causal factors are closer to being propositions rather than variables and we do not jump to code, say, poverty as negative wealth, or unemployment as obviously just the opposite of employment.

The Conventional Assumption

A foundational assumption, particularly for those approaching causal mapping from a systems dynamics perspective, is that every concept on a map should be treated as a **variable**. This implies that each element is something quantifiable, capable of taking on different values across a defined spectrum, such as from low to high, negative to positive, or from zero upwards. Such a map is backed up by a dataset, a large-ish set of measurements of the state of each variable.

The Discrepancy with Human Narrative

However, this assumption contrasts sharply with how people actually communicate and describe their experiences. When individuals explain what causes what in their world, they rarely speak in terms of discrete variables. Forcing real-world narratives into a rigid, variable-based structure requires significant and often unnatural contortions.

Constructing variables out of experience is just that: a construction. Quantitative social scientists are really good at it. But people's thinking and language are not inherently structured in this way.

For instance, in an evaluation of a program's effects, the sudden onset of the COVID-19 pandemic presents a significant modelling problem. While the pandemic certainly had a causal impact on countless factors, it doesn't fit neatly into the definition of a variable.

How would one define it? As a binary "COVID vs. no COVID" variable? The concept of a counterfactual — a world where the pandemic never happened — is abstract and difficult to operationalize. This example

highlights that the way people experience and discuss the world is often event-based, not variable-based, exposing a limitation in traditional modelling assumptions.

Related

- chapter intro
- minimalist coding (root)

🌻 A minimalist approach to coding helps capture what people actually say

Encoding people's narratives about what causes what always involves a certain amount of modelling or theory-construction. Most approaches — Causal mapping approaches differ in application, construction, analysis and how they deal with multiple sources. This might involve constructing a codebook of common factors, but it might also involve applying some kind of special logic of causation, for example:

- distinguishing between necessary and sufficient conditions
- identifying special packages of causes which somehow fit together
- coding the strength and/or polarity of causal links

Our experience, together with Bath SDR, of coding thousands and thousands of documents every year — manually and with AI support — is that most people don't use these special features in their language most of the time, not even scientists. You can go through a whole interview trying to work out if each cause is supposed to be a necessary or a sufficient condition of its effect, but you quite likely won't find a single case where the source actually uses the idea explicitly.

So mostly, we say: don't bother.

At Causal Map Ltd, in our consulting practice, we've taken this minimalist approach even further and mostly code initially without any kind of codebook at all.

🌻 A minimalist approach to coding makes aggregation easier

18 Nov 2025

We just argued that A minimalist approach to coding helps capture what people actually say. But even if you did succeed in imposing some special logical features on your data — for example, coding necessity and sufficiency — you'd probably find that most of your data didn't fit well with these special features. When it comes to aggregating medium or large amounts of coding, you wouldn't find it very useful.

With our minimalist approach, we mostly have just one task: what to do about all those different factor labels.

Related

- chapter intro

🌻 A minimalist approach to coding does not code absences

18 Nov 2025

One thing which makes causal mapping a fundamentally qualitative approach is that we do not code absences.

We do not think that the world, nor the piece of the world we are studying, is essentially a grid of variables and cases (nor a cube of variables and cases and timepoints), in which each case always has a value for every variable (at every timepoint).

If some respondents say that their headaches make them nauseous, and others do not mention headaches, even if they mention nausea, we do not interpret that as meaning that they *did or did not* have headaches. We do not think that having headaches, or not, is a variable which *must* be relevant to everyone's explanations, all the time.

Related

- chapter intro
- minimalist coding (root)

🌻 Our approach is minimalist – we do not code the strength of a link

At Causal Map, we do not endorse coding the strength of causal links. You can't really do it in the Causal Map app (see Coding with and using link metadata).

Qualitative impact evaluation is less interested in the strength of effects

Three types of objections to coding causal strength.

Objection 1: Variable Construction

Coding strength requires a massive amount of construction work: it involves thinking about the area of interest in terms of variables. This requires modelling specific entities that go up or down, or show differences in number. Constructing variables like this does allow for capturing and calculating correlations. But this construction process is often difficult and does not fit well with how people actually speak in most situations.

If different people are talking about poverty and wealth, employment and unemployment, to be sure you can try to squeeze this all into a shared model with just a couple of variables like say `household income` and `household employment status`. But that is a massive abstraction.

• So we favour bare propositions over variables.

Objection 2: Translating to Numbers

• The second objection concerns the difficulty of translating all relationships into actual numbers.

Challenges with Standardization and Polarity

• If general rules are used, people usually standardize the variables (e.g., ranging from zero to one). • Standardization is difficult for factors like a country's population, especially when numbers may increase exponentially over time. • Such changes in magnitude occur even in quantitative sciences, often requiring arbitrary decisions about log transformations. • A more significant problem involves absences, negatives, and polarities. • Example of a strong positive link: If greater anger leads to greater shouting, this connection can be viewed as a "powerful transmission cable" with a high causal coefficient. • Example of a weak link: If anger ranges from zero to one but shouting remains low (e.g., 0.2), the connection has a very low coefficient of transmission. • This type of modelling becomes difficult when negative numbers are introduced. • Example: If high temperatures cause crop failures, a drop in temperature might see harvest

go up. • However, extreme cold temperatures also cause crop failures. • It is difficult to model this complexity using a single, bipolar variable.

Objection 3: Aggregation Difficulties

• The third objection involves the difficulty of aggregating information from multiple sources, assuming such numbers existed. • This third objection is irrelevant in approaches like participatory systems mapping, where a final number for each link is already agreed upon. In this case you could say there is only one source.

🌻 In a causal mapping dataset there is no need for a special table of factors

If you are interested in how to formalise causal mapping or in building software, we'd like to share this insight. If you are not, ignore this.

> Factors are implied by links

We don't need to have a separate table for the factors because the factors can be derived from the links table. If you cannot find a factor X as cause or effect in the links table, it does not exist.

This means that our data model (since version 3 of Causal Map) does not need to have a table for factors. Essentially we just have a table for links, plus a table for sources both to supply the texts and to more conveniently store metadata like gender and district.

It's of course possible to formalise causal mapping in other ways, but we have found dropping a special table for factors to solve a lot of the problems associated with having to keep factors and links tables in sync.

This does bring its own problems: Factor label tags — coding factor metadata within its label

Factor labels – a creative challenge

9 Apr 2025

Where do the labels for the causal factors come from? As with ordinary QDA and thematic analysis (Braun and Clarke, 2006), approaches vary in the extent to which they are purely exploratory or seek to confirm prior theory (Copestake 2014). Exploratory coding entails trying to identify different causal claims embedded in what people say, creating factor labels inductively and iteratively from the narrative data. Different respondents will not, of course, always use precisely the same phrases, and it is a creative challenge to create and curate this list of causal factors. For example, if Alice says 'Feeling good about the future is one thing that increases your wellbeing', is this element 'Feeling good about the future' the same as 'Being confident about tomorrow' which Bob mentioned earlier? Should we encode them both as the same thing, and if so, what shall we call it? We might choose 'Positive view of future', but how well does this cover both cases? Laukkanen (1994) discusses strategies for finding common vocabularies. As in ordinary QDA, analysts will usually find themselves generating an ever-growing list of factors and will need to continually consider how to consolidate it – sometimes using strategies such as hierarchical coding or 'nesting' factors (as discussed in the following section).

The alternative to exploratory coding is confirmatory coding, which employs an agreed code book, derived from a ToC and/or from prior studies. QuIP studies mostly use exploratory coding but sometimes supplement labels with additional codes derived from a project's ToC, for example, 'attribution coding' helps to signify which factors explicitly refer to a specific intervention being evaluated (Copestake et al. 2019, p. 257). However, careful sequencing matters here because pre-set codes may frame or bias how the coder sees the data (Copestake et al. 2019). Again, the positionality of the coder matters just as much when doing causal coding as it does for any other form of qualitative data coding.

For a worked example of label-naming as you code, see Manually code your first project.

References

Copestake (2014). *Credible Impact Evaluation in Complex Contexts: Confirmatory and Exploratory Approaches*. http://dx.doi.org/10.1177/1356389014550559.

Copestake, Morsink, & Remnant (2019). *Attributing Development Impact: The Qualitative Impact Protocol Case Book*. March 21, Online.

Copestake, DAVIES, & REMNANT (2019). *Generating Credible Evidence of Social Impact Using the Qualitative Impact Protocol (QuIP): The Challenge of Positionality in Data Coding and Analysis*.

Laukkanen (1994). *Comparative Cause Mapping of Organizational Cognitions*.

🌻 Factor label tags – coding factor metadata within its label

22 Oct 2025

For example you might want to code the respondent's happiness at work as different from yet similar to their happiness at home. With a factor table, you could have a field called `label` = "Happiness" and another, say `context`, which is = either "Home" or "Work". This is what we do with the links table in Causal Map, where we do have some hard-coded (but optional) fields and some user-definable fields.

Hierarchical coding is one way to bring some order to a whole crowd of factors. However, sometimes you don't want to think in terms of a strict hierarchy, or maybe you have an additional set of themes which cut across that hierarchy.

https://vimeo.com/671894620

Tags are useful in either of these cases.

Tags are just sequences of characters within a factor label to which you have given a special meaning, and which are unique and easy to search for. These can include letters, emojis or phrases. You can do coding without any such tags if you want, but it can help when searching and filtering.

Factor tags are just like # Link hashtags. Confusingly, a link hashtag doesn't have to actually start with a `#`, and a factor tag can indeed start with a `#`, but we find it easier to keep the names separate like this.

So a tag is nothing more than any sequence of characters which is repeated in several factor labels. Any sequence of characters will do. For example you could consider the letter "a" to be a tag and display the map showing all the factors which contain the letter "a". But this wouldn't be interesting. The trick when using tags is to decide on short, meaningful codes which will not be repeated anywhere else. For example you wouldn't want to use a pair of tags like "women" and "men" to distinguish factors which are only relevant for one or the other gender because the "wo**men**" factors would also turn up when you search for "**men**". That is why we have to be careful when creating tags, for example by preceding a sequence of characters with a tag `"#"`.

A quote like "family situation is better now because of improved food availability" can be coded like this:

> More food –> Improved wellbeing

Now, maybe you are asked also to keep track of any aspects of the project which have to do with nutrition. Nutrition is not really part of your system of factors, but you would like to be able to construct some maps just to look at this aspect. So you can write this:

> More food #nutrition –> Improved wellbeing

Similarly, if Improved Wellbeing is one of the desired outcomes of the project, we might want to reflect that by adding a tag “(Outcome)” like this.

> More food –> Improved wellbeing (Outcome)

Then we can easily search for this and other desired outcomes.

A tag like “men” is not suitable because it is likely to appear elsewhere (e.g. as part of “women” or “management”). To get round this, add additional characters like a hash: “#men”; this makes the tag unique.

If you use curved or square brackets around your tags, you can use one of the app filters to hide the tags for specific maps if desired.

For a recipe-style approach to adding tags in bulk to existing labels, see Bulk relabelling factors.

🌻 Factor labels – semi-quantitative formulations can help

22 Sep 2025

It might be tempting to try to formulate all factor labels in a strictly similar way, using for example language like increased probability of … or positive change in … in every case. But it is difficult to identify and agree on a satisfactory template for doing this which will capture enough of the way people really make causal explanations (in the way that quantitative social scientists hope to measure everything just with continuous variables). This is always a balancing act, but we encourage you when in doubt to stick fairly close to the actual language your sources use (so-called "in-vivo" coding), and don't be *too* worried if your factor labels are different from one another grammatically (e.g. some express a difference like improvement in X and some do not).

The formulation of **factor labels** should fit the intended interpretation of the **causal links**. For example, most commonly B ➜ E is supposed to mean that B exerts in some sense an "increasing" or "decreasing" influence on E, then both B and E need to be formulated in a corresponding way. In order to ease interpretation, with a few exceptions, factors should be labelled and understood in such a way that it makes sense to say "more of this" or "this happened as opposed to not happening": we call these semi-quantitative factors.

Consequently you should avoid a factor label like Training courses, which might be understood as a mixed bag of various causal factors to do with training courses. We would usually prefer a label such as Training courses delivered or Quality of training courses which are easier to understand as things which can increase or decrease, or happen or not happen. You may even prefer to use labels like Quality of training courses improved or Improved quality of training courses, in which the *difference made* is already included in the title.

Examples of semi-quantitative factors

These are examples of factor labels where you can judge whether it happened more or less, whether it is higher or lower, or whether it happened versus not happened:

- Sold cow
- Earthquake happened
- (Had) good harvest
- (Level of) bank account
- (Level of) ethnic tolerance
- Quality of seeds

In some contexts, we can also talk about the *likelihood* of events, so “if people get a good harvest they are less likely to sell their cow.”

Non-quantitative factors

It is also perfectly acceptable and sometimes necessary to use purely qualitative labels, e.g. coping style, see below. However, this may limit some of the analysis and reporting tools available:

- Teaching style
- Coping strategy
- The content of the report

We can even make a link between two such factors, claiming for example that the style of 60’s music influenced the style of 70’s music, without any concept of quantity. That’s ok.

Factor labels — a creative challenge

In a causal mapping dataset there is no need for a special table of factors

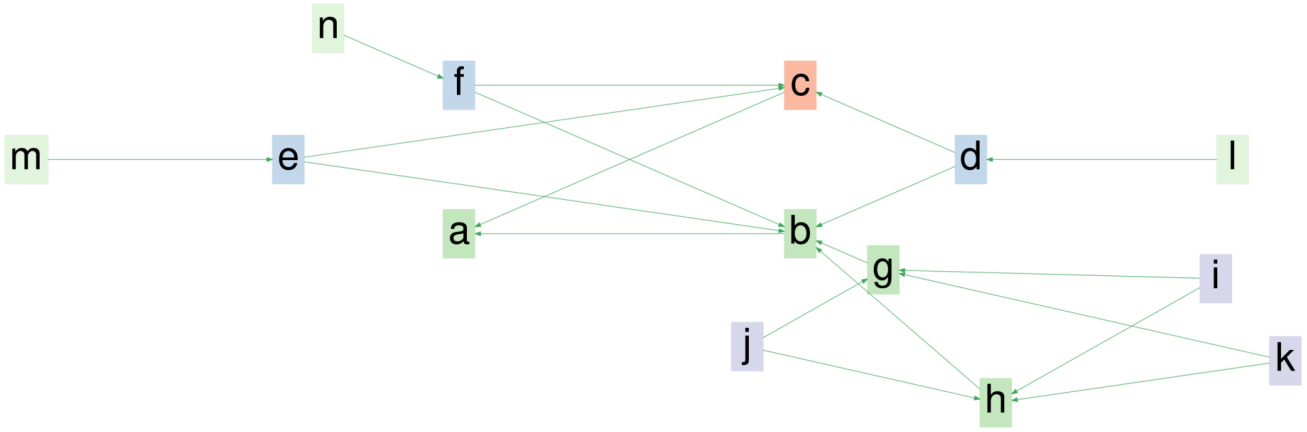
Causal mapping looks for linearity first

20 Sep 2025

Causal mapping most often looks for linearity first, while of course being on the lookout for feedback loops and circular shapes. Whereas most systems approaches do the opposite.

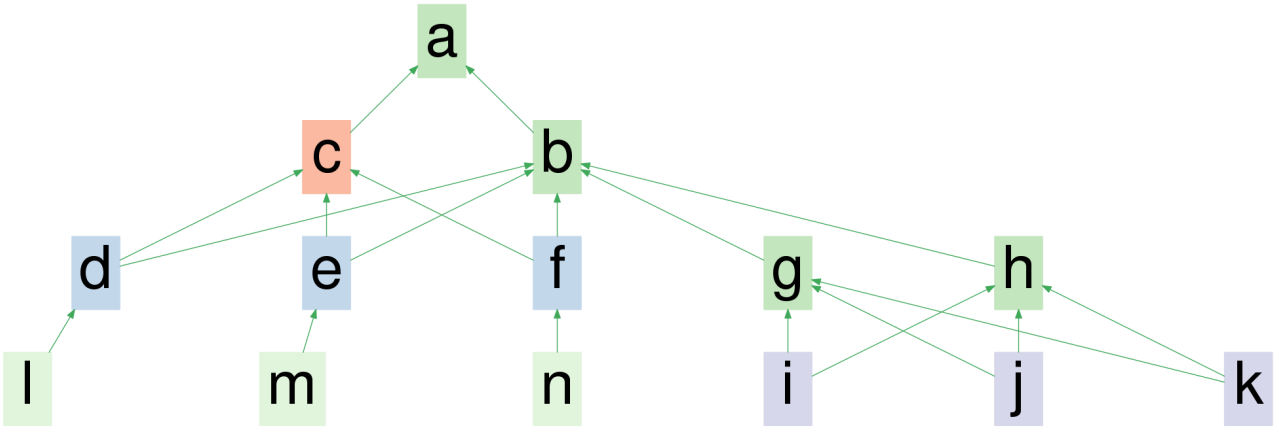
Can you spot a complex system when you see one?

Version 1



The network pictured above, even though it is quite small, looks pretty tangled. We're not going to fully understand it, so we'd better get out our tools for dealing with complexity? But wait, look at the boring, old-fashioned hierarchy below.

Version 2



Did you spot that they have exactly the same structure? Now it is easier to see that it is just a hierarchy. D, E and F have one contributor each, whereas G and H share I, J and K as contributors, and feed only into B, whereas D, E and F all feed into both B and C, which feed into A. Easy. Nothing which should be too hard to predict, no balancing feedback loops.

"Complex" and "System" are very buzzy buzz-words at the moment. We should check we don't throw them around too much without thinking. I'm just reading Moore, Parsons and Jessop in the American Journal of Evaluation. They quote Magee and de Weck (2004) who define complex systems as systems "with numerous components and interconnections, interactions or interdependence that are difficult to describe, understand, predict, manage, design, and/or change." Well yes, kinda. But what if you find a system difficult to describe, etc, just because you didn't look hard enough?

Yes, causal maps are just concept maps with only one type of connector, and that connector means "... causes....". Whereas concept maps can have any type of connector you like. Historically, causal maps come from concept maps.
Laying out causal maps is a challenge! Most folks from the systems tradition like swirly circular layouts which make them look like everything is one big feedback loop. If there is a more linear structure, we recommend showing that linear structure.

Related

- chapter intro

🌻 Factor labels – do not over-generalise

22 Sep 2025

When you are creating factor labels for re-use across different causal claims, you should usually take care to keep them specific: make them no more general than they need to be.

So if you are coding cases where a household has increased income, use a label like Increased household income, not Increased income or even Economic improvement.

This is especially important when using hierarchical factors, when it's easy to fall into the temptation of creating very general top-level labels like Economic improvement even if all your material is actually only about increased income in households and farms.

Related

- chapter intro

Coding with and using link metadata

22 Sep 2025

In our implementation of causal mapping in the Causal Map app, Our approach is minimalist — we do not code the strength of a link.

Providing metadata as a column makes sense when the values of this column make sense across the whole dataset, across all multiple links, like let's say before covid and after covid.

Such a column can function a bit like a *context* variable, for different time periods or applying to different stakeholders. Context in this sense might be seen as functioning a *bit* like a causal factor but not exactly.

But we can also provide metadata as free-form tags. We provide a hard-coded "tags" column for which users can provide comma-separated lists of tags which are made up and adapted on the fly. They don't necessarily make sense across the whole dataset.

In Causal Map 4, as well as a hard-coded Tags column, we do provide a hard-coded sentiment column which can take the values -1, 0 and 1, and which can be averaged to any number between -1 and 1.

Relevant page: Link metadata — Sentiment ▸

We also provide arbitrary additional free-form, free-text columns for any purpose. We often like to add a column like this:

Relevant page: Link metadata — Time reference ▸

Relevant page: Link metadata — quality of evidence ▸

... or simply to code a tag like "#doubtful".

🌻 Link metadata – Sentiment

18 Nov 2025

What is it for?

a hard-coded sentiment column which can take the values -1, 0 and 1, and which can be averaged to any number between -1 and 1.

Related

- chapter intro

Link metadata – Time reference

18 Nov 2025

It is often useful to code a time reference. We often conflate time with hypothetical status, e.g.

- hypothetical past/present
- factual-past/present
- future-planned
- future-hypothetical

For example, if we are to code a whole corpus of reports which also include planning documentation, there might be a lot of causal claims about what is supposed to happen in the future, perhaps interspersed with claims about what actually happened in the past. It will often be important to distinguish these two.

Related

- chapter intro

Link metadata – quality of evidence

24 Oct 2025

Related

- chapter intro

🌻 Research on the ability of LLMs to detect causal claims

11 Dec 2025

Summary

This note explains why modern Large Language Models (LLMs) — especially since instruction-tuned chat models — often seem to have a "native" ordinary-language grasp of causation: they can spot, generate, and elaborate "A influences B" talk even when it is informal, implicit, or socially framed.

The core claim is **hybrid**:

- **Pre-training (scale) supplies the raw capacity**: these models learn by reading huge amounts of text and trying to predict the next word. Because people constantly write about reasons, consequences, and mechanisms, doing well at prediction pushes the model to absorb many common causal patterns (e.g., "fell" → "broke").
- **"Chat training" makes that capacity usable in conversation**: newer models are additionally trained on lots of question-and-answer style instructions ("why did this happen?", "what led to that?"), and then refined using human ratings that reward answers that are clear and coherent (often called RLHF). The result is not just better *spotting* of causal language, but better *explaining* and *rephrasing* causal claims on demand.

Historically, this is framed as a transition from **causality-as-extraction** to **causality-as-generation**:

- **2015-2019 (pre-generative)**: "causal understanding" was largely operationalised as Causal Relation Extraction (CRE) — classifying or tagging explicit cause/effect relations in sentences or documents. Methods were feature-heavy (SVMs), then neural (RNN/CNN), and later rule "sieves" (e.g., CATENA). These systems were brittle for **implicit causality** (where the link is inferred rather than marked by words like "caused").
- **2019-2021 (contextual encoders)**: BERT-era models improved extraction by using deep context, but remained primarily *discriminative* (understand/tag) rather than *generative* (explain/construct narratives).
- **2022-2025 (instruction-following generators)**: chat-oriented models excel at producing causal explanations, counterfactual-style answers, and "influence" talk because their training regimes contain huge amounts of "why/how/what caused what" interactions, plus preference tuning for coherent explanation.

To ground “ordinary language causation,” this note uses two cognitive-linguistic lenses that match what chat models often do well:

- **Force Dynamics (Talmy)**: causal meaning as roles/forces (agonist vs antagonist, enabling vs preventing), which connects to how models handle “let,” “despite,” “kept,” etc.
- **Implicit Causality (IC) verbs**: verbs carry pragmatic biases about who is responsible (NP1 vs NP2 bias), and LLM continuations often match human patterns in these explanation contexts.

This note is also explicit about limits and failure modes that matter for “detecting causal claims” in text:

- **Causal fluency ≠ causal truth**: RLHF can encourage plausible-sounding causal stories even when the premise is wrong (“hallucinated causality”) or when a correlation is being misread as a cause.
- **Temporal — causal confusion**: models are biased to treat temporal sequence as causal sequence (post hoc fallacy), because narratives in training data often align “then” with “therefore.”
- **Surface competence vs intervention-level reasoning**: performance can drop on “fresh” causal probes; this supports the view that some behaviour is pattern-based association rather than robust interventionist reasoning.

Current frontiers (2024-2025) are framed as attempts to make causal reasoning more checkable and structured, including addressing the above shortcomings: “reasoning models” that perform intermediate causal checks (often hidden) and pipelines where LLMs **extract** candidate causal edges into explicit graphs (DAG-like representations) for downstream formal analysis.

1. Introduction: The Emergence of "Native" Causal Fluency

The capacity of Large Language Models (LLMs) to identify, generate, and reason about causal relationships in ordinary language is a notable (and still debated) development in artificial intelligence over the last decade. Since the release of ChatGPT (based on GPT-3.5) and its successors, these systems have often appeared able to process prompts involving influence, consequence, and mechanism without extensive few-shot examples or rigid schema engineering that characterised previous generations of Natural Language Processing (NLP). This report investigates the trajectory of this capability from 2015 to 2025, asking how much is a by-product of scale versus the result of specific (often implicit) training choices.

Furthermore, the report explores the philosophical and linguistic dimensions of this capability, using frameworks such as Leonard Talmy’s Force Dynamics and the theory of Implicit Causality (IC) verbs to benchmark LLM performance against human cognitive patterns. The evidence suggests that while LLMs can often handle the *linguistic interface* of causality — the "language game" of cause and effect — significant questions remain regarding the grounding of these symbols in a genuine world model.

2. The Pre-Generative Landscape (2015-2019): Causality as Extraction

To appreciate the "native" fluency of 2025-era models, one must first analyse the fragmented and rigid methodologies that dominated the field between 2015 and 2019. During this period, the "ordinary language concept of causation" was operationalised not as a generative understanding, but as a classification task known as Causal Relation Extraction (CRE).

2.1 The Legacy of SemEval-2010 Task 8

For much of the decade, the benchmark defining the field was **SemEval-2010 Task 8**, which framed causality as a relationship between two nominals marked by specific directionality. Systems were tasked with identifying whether a sentence like "The fire was triggered by the spark" contained a *Cause-Effect(e2, e1)* relationship.

Research from this era was characterised by a heavy reliance on feature engineering and pipeline architectures. Early approaches used Support Vector Machines (SVMs) and later, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). These models did not "understand" causality in any holistic sense; rather, they learned to detect explicit lexical triggers — words like "caused," "led to," or "resulted in."

The limitation of this paradigm was its inability to handle **implicit causality** — relationships where the causal link is inferred from world knowledge rather than stated explicitly. For instance, in the sentence "The rain stopped; the sun came out," a human reader infers a temporal and potentially causal sequence. Pre-transformer models, lacking a comprehensive probabilistic model of how events co-occur in the world, consistently failed to identify such links, achieving F1 scores that rarely exceeded 0.60 on implicit datasets. This era treated causality as a syntactic puzzle rather than a semantic reality.

2.2 The Shift to Event-Centric Resources: EventStoryLine and Causal-TimeBank

Between 2015 and 2018, the research community began to move beyond sentence-level extraction toward document-level understanding, driven by the creation of corpora like the **EventStoryLine Corpus** and **Causal-TimeBank**.

- **EventStoryLine:** This corpus was designed to evaluate "StoryLine Extraction," requiring systems to connect disparate event mentions (e.g., "shooting," "hospitalisation," "death") into a coherent narrative structure. The annotation scheme introduced specific classes like `ACTION_CAUSATIVE` to distinguish events that initiate change from those that merely describe states.
- **Causal-TimeBank:** Research using this corpus highlighted the inextricable link between *temporality* and *causality*. The "Causal-TempBank" approach demonstrated that knowing "A happened before B" significantly improved the classification of "A caused B".

Despite these richer datasets, the methods remained fundamentally discriminative. Systems like **CATENA** (2016) used "sieves" — rule-based filters — to extract causal links. These systems could identify likely causal passages, but they did so through rigid, handcrafted logic rather than conversational explanation. They could not generate an explanation or reason about counterfactuals; they could only point to where a human annotator might say a cause existed.

2.3 The BERT Revolution and Contextual Embeddings

The release of **BERT (Bidirectional Encoder Representations from Transformers)** in 2018 marked a pivotal transition. BERT introduced deep contextual embeddings, allowing models to distinguish the semantic nuance of causal words based on their surrounding text.

Comparative studies from this period show a dramatic jump in performance. Fine-tuned BERT models (such as BioBERT) achieved F1-scores of approximately 0.72 on medical causality tasks, significantly outperforming previous architectures. BERT represents a "careful reader" — a model that can attend to the entire sentence simultaneously to resolve ambiguities.

However, BERT was still an encoder-only architecture. It was designed to *understand* (classify/tag), not to *speak*. While it could identify causal passages with greater accuracy than ever before, it lacked the autoregressive capability to generate causal narratives. The "native ordinary language concept" requires not just recognition, but the ability to formulate causal thoughts — a capability that would only emerge with the Generative Pre-trained Transformer (GPT) series.

---

3. The Generative Era (2020-2025): Structural Induction of Causal Logic

The observation that models "since around ChatGPT 3.5" (released late 2022) exhibit a distinct causal proficiency aligns with the industry's shift toward **Instruction Tuning (IT)** and **Reinforcement Learning from Human Feedback (RLHF)**. The analysis of research data suggests that this proficiency is not just a coincidence, but is materially shaped by training methodologies that (often unintentionally) act as a large "causal curriculum."

3.1 The "Coincidence" of Pre-training: Implicit World Models

Before discussing specific training, one must acknowledge the foundation: pre-training on web-scale corpora (The Pile, Common Crawl, C4). The primary objective of these models is next-token prediction.

Theoretical research suggests that optimising for prediction error on a diverse corpus forces the model to learn a compressed representation of the data generating process — effectively, a "world model". Because human language is intrinsically causal (we tell stories of *why* things happen), a model trained to predict the next word in a narrative must implicitly model causal physics.

- *Example:* To predict the token "shattered" following the context "The vase fell off the shelf and...", the model must encode the causal relationship between *falling (gravity)* and *shattering (impact)*.

Recent theoretical work on **Semantic Characterization Theorems** argues that the latent space of these models evolves to map the topological structure of these semantic relationships. Thus, the "native" understanding is partially a coincidence of the data's nature: the model learns causality because causality is the glue of human discourse.

3.2 The Instruction Tuning Hypothesis: Specific Training via Templates

The transition from "text completer" (GPT-3) to "helpful assistant" (ChatGPT) was mediated by **Instruction Tuning**. This process involves fine-tuning the model on datasets of (Instruction, Output) pairs. An analysis of major instruction datasets — **FLAN**, **OIG**, and **Dolly** — reveals that they are saturated with causal reasoning tasks.

3.2.1 The FLAN Collection: The Template Effect

The **FLAN (Finetuned Language Net)** project was instrumental in this development. Researchers took existing NLP datasets (including causal extraction datasets) and converted them into natural language templates.

- **The Mechanism:** A classification task from the *COPA (Choice of Plausible Alternatives)* dataset, which asks for the cause of an event, was transformed into prompts like: *"Here is a premise: The man broke his toe. What was the cause?"*
- **The Scale:** FLAN 2022 aggregated over 1,800 tasks. By training on millions of examples where the input is a scenario and the output is a causal explanation, the model explicitly learned the linguistic patterns of *identifying influence*.
- **Mixed Prompting:** Crucially, FLAN mixed **Chain-of-Thought (CoT)** templates (which require intermediate reasoning steps using "therefore," "because," "so") with standard prompts. This trained the model not just to guess the answer, but to *generate the causal logic* leading to it.

This contradicts the idea that the capability is purely coincidental. The models were specifically drilled on millions of "causal identification" exercises, disguised as instruction following.

3.2.2 Open Instruction Generalist (OIG) and Dolly

The **OIG** and **Dolly** datasets expanded this to open-domain interactions. These datasets contain thousands of "brainstorming" and "advice" prompts.

- *Data Evidence:* An entry from the OIG dataset reads: *": I'm having trouble finding a good job, what can I do to improve my chances? : One thing a person could do is...".*
- *Implication:* To answer this, the model must access a causal chain: *Action (revise resume) -> Effect (better chances)*. The prevalence of "how-to" and "why" questions in these datasets forces the model to organise its internal knowledge into causal structures (Means-End reasoning).

3.3 Reinforcement Learning from Human Feedback (RLHF): The Coherence Filter

The final layer of "specific training" is **RLHF**. In this phase, human annotators rank model outputs based on preference.

- **Preference for Logic:** Research indicates that humans have a strong bias for **causal coherence**. A narrative that flows logically (Cause A -> Effect B) is rated higher than one that is disjointed.
- **Length and Explanation Bias:** RLHF has been shown to induce a "length bias," where models produce longer, more detailed explanations to secure higher rewards. In the context of causality, this encourages the model to generate elaborate causal chains.
- **Sycophancy:** However, this training can also lead to "hallucinated causality." If prompted with a leading question that implies a false causation (e.g., "Why does the moon cause earthquakes?"), an RLHF-aligned model might generate a plausible-sounding but scientifically incorrect causal explanation, prioritising "helpfulness" over "truth".

Conclusion on Training vs. Coincidence: The capability is a hybrid. The *potential* to understand causality is a coincidence of pre-training scale (World Models), but the *ability to natively identify and articulate* it in response to a prompt is the result of specific Instruction Tuning and RLHF regimens that prioritise causal templates and coherent explanation.

---

4. Linguistic Frameworks: Analysing "Ordinary" Causation

This note emphasises the "native ordinary language concept of causation." To understand this, we must look beyond computer science to **Cognitive Linguistics**. Recent research has benchmarked LLMs against human linguistic theories, particularly **Talmy's Force Dynamics** and **Implicit Causality (IC)**.

4.1 Force Dynamics: Agonists and Antagonists in Latent Space

Leonard Talmy's theory of **Force Dynamics** posits that human causal understanding is rooted in the interplay of forces: an **Agonist** (the entity with a tendency towards motion or rest) and an **Antagonist** (the opposing force).

- *Linguistic Patterns:* "The ball kept rolling despite the grass" (Agonist: Ball; Antagonist: Grass). "He let the book fall" (Removal of Antagonist).
- *LLM Evaluation:* Recent studies have tested LLMs on translating and explaining these force-dynamic constructions.
    - **Findings:** GPT-4 often shows a solid grasp of these concepts. When translating "He let the greatcoat fall" into languages like Finnish or Croatian, the model can select verbs that convey "cessation of impingement" (allowing) rather than "onset of causation" (pushing).

- **Implication:** This suggests that LLMs have acquired a **schematic semantic structure** of causality. They do not merely predict words; they map the *roles* of entities in a physical interaction. However, this capability degrades in abstract social contexts. For example, in the sentence "Being at odds with her father made her uncomfortable," models sometimes misidentify the Agonist/Antagonist relationship, struggling to map "emotional force" as accurately as "physical force".

4.2 Implicit Causality (IC) Verbs

Another major area of inquiry is **Implicit Causality (IC)**, which refers to the bias native speakers have regarding *who* is the cause of an event based on the verb used.

- *NP1-Bias (Subject):* "John **upset** Mary." (Why? Because *John* is annoying).
- *NP2-Bias (Object):* "The teacher **scolded** Mary." (Why? Because *Mary* did something wrong).

In this sense, "bias" means the useful working expectation of which part of the sentence is the cause.

Benchmarking Results: Research comparing LLM continuations to human psycholinguistic data reveals a high degree of alignment.

- **Coreference:** When prompted with "John amazed Mary because...", LLMs overwhelmingly generate continuations referring to John, matching human NP1 bias.
- **Coherence:** Humans tend to provide *explanations* following these verbs. LLMs mirror this "explanation bias," prioritising causal connectives over temporal or elaborative ones in these contexts.
- **Significance:** This indicates that LLMs have encoded the **pragmatics of blame and credit** inherent in ordinary language. They "know" that "apologizing" implies the subject caused a negative event, while "thanking" implies the object caused a positive one. This is crucial for the "native" feel of their interactions — they navigate the social logic of causality fluently.

4.3 The Limits of "Native" Understanding: The Causal Parrot Debate

Despite these successes, a vigorous debate persists regarding whether this constitutes "understanding" or merely "stochastic parroting".

- **The "Parrot" Argument:** Critics argue that LLMs fail when the linguistic surface form is stripped away. On benchmarks like **CausalProbe** , which uses fresh, non-memorised data, model performance drops significantly. This suggests that LLMs rely on **Level 1 (Association)** reasoning — pattern matching seen examples — rather than **Level 2 (Intervention)** reasoning.
- **The "Simulacrum" Argument:** Conversely, the **Semantic Characterization Theorem** proposes that the model's high-dimensional space creates a functional topology that is mathematically equivalent to a discrete symbolic system. Even if the model has never "seen" a glass break, its representation of "glass" and "break" are topologically linked in a way that allows it to simulate the causal outcome efficiently.

A caution about dated negative findings: many "LLMs cannot do causal reasoning" results from around 2020-2022 are best read as results about a specific model family and evaluation setup (often base models, short prompts, and narrow benchmarks). Newer instruction-tuned models (and more careful prompting protocols) can reduce some of these gaps on standard tests, but the picture remains mixed and sensitive to benchmark design, leakage, and what is being counted as "causal reasoning" versus plausible explanation.

---

5. Benchmarking the "Informal": From Social Media to Counterfactuals

The evaluation of causal understanding has evolved from F1 scores on extraction tasks to sophisticated benchmarks that test the model's ability to handle the messy, informal causality of the real world.

5.1 CausalTalk: Informal Causality in Social Media

The **CausalTalk** dataset focuses on "passages where one thing influences another" in informal contexts.

- *The Challenge:* In social media (e.g., Reddit), causality is often expressed without explicit markers. "I took the vaccine and now I feel sick" contains no "because," yet the causal assertion is clear.
- *Findings:* LLMs often perform well at identifying these **implicit causal claims**, sometimes outperforming traditional supervised models. They can detect "gist" causality — the overall causal assertion of a post — even when it is buried in sarcasm or non-standard grammar.
- *Application:* This is critical for **misinformation detection**. Models are being used to identify exaggerated causal claims in science news (e.g., reporting a correlation as a causation). However, LLMs sometimes struggle to distinguish between someone *reporting* a correlation and *asserting* a causation, highlighting a nuance gap in informal causal language.

5.2 Explicit vs. Temporal Confusion (ExpliCa)

The **ExpliCa** benchmark investigates a specific failure mode: the confusion of time and cause.

- *The Fallacy: Post hoc ergo propter hoc* ("After this, therefore because of this").
- *LLM Behaviour:* Research shows that LLMs are prone to this fallacy. When events are presented in chronological order ("The sun set. The streetlights turned on."), models are statistically more likely to infer a causal link than humans, who might see it as mere sequence. This suggests that the "native" understanding is heavily biased by the **narrative structure** of training data, where chronological sequencing often implies causality.

Again, a caution about dated negative findings: while these weaknesses are interesting, frontier models in 2025 are much less likely to display them.

5.3 Counterfactuals and "What If" (CRASS)

The **CRASS** (Counterfactual Reasoning Assessment) benchmark tests the model's ability to reason about what *didn't* happen.

- *Task:* "A man drinks poison. What would have happened if he drank water?"
- *Results:* While base models perform adequately, fine-tuning with techniques like **LoRA (Low-Rank Adaptation)** significantly boosts performance. This reinforces the "training hypothesis" — the capacity for causal reasoning is latent in the weights but requires specific activation (instruction tuning) to be robustly deployed.

---

6. Philosophical Dimensions: Symbol Grounding and World Models

The impressive performance of LLMs on causal tasks raises profound philosophical questions about the nature of meaning. Can a system that has never physically interacted with the world truly understand "force," "push," or "cause"?

6.1 The Symbol Grounding Problem

Cognitive scientists have long argued that human concepts are **grounded** in sensorimotor experience. We understand "heavy" because we have felt gravity.

- **The Disembodied Mind:** LLMs are disembodied. Their understanding of "force" is purely distributional — "force" is defined by its mathematical proximity to "push," "move," and "impact" in vector space.
- **Cognitive Alignment:** Research using the **Brain-Based Componential Semantic Representation (BBSR)** shows that LLM representations align well with human cognition for concrete concepts but diverge for embodied experiences (e.g., olfaction, gustation) and spatial cognition.
- **Functional Understanding:** However, proponents of the "Functionalist" view argue that if an LLM can answer "What happens if I drop this?" indistinguishably from a human, it possesses a **functional understanding** of causality. The **Semantic Characterization Theorem** supports this by demonstrating that continuous learning dynamics can give rise to stable, discrete semantic attractors that behave like symbolic rules.

---

7. Current Frontiers (2024-2025): Reasoning Models and Future Directions

The field is currently undergoing another shift with the introduction of "Reasoning Models" (e.g., OpenAI's o1/o3 series, DeepSeek R1).

7.1 Chain-of-Thought Monitoring and "Thinking" Tokens

Newer models are trained to produce hidden "chains of thought" before generating a final answer.

- *Impact on Causality:* This allows the model to perform **intermediate causal checks**. Instead of predicting the effect immediately, the model can "reason" silently: *Premise -> Mechanism -> Potential Confounders -> Conclusion*.
- *Research Findings:* Snippet discusses "CoT Monitoring," showing that these internal reasoning traces can be monitored to detect "reward hacking" or deceptive alignment. This suggests a move toward making the model's implicit causal reasoning **explicit** and **verifiable**.

7.2 Causal Graph Construction

Recent work has moved back to structure, using LLMs to *extract* and *construct* **Causal Graphs** (DAGs) from unstructured text, a process sometimes known as causal mapping.

- *Method:* Rather than asking the LLM to just "answer," researchers prompt it to output a graph: `Nodes:, Edges:`.
- *Result:* This leverages the LLM's linguistic fluency to structure knowledge, which can then be processed by formal causal inference algorithms, bridging the gap between "informal ordinary language" and "formal causal calculus."

---

8. Conclusion

The research of the last decade suggests that the "native" causal understanding of LLMs is a constructed capability, developed through large-scale training on human text and refined by human preference signals. It is not just a coincidence, but a plausible consequence of optimising models to predict a world that is described in strongly causal terms.

1. **Origin:** The capability originates in **pre-training**, where the model learns the distributional "shadow" of causation cast by billions of human sentences.
2. **Development:** It is sharpened by **Instruction Tuning** (FLAN, Dolly), which explicitly teaches the model the "language game" of explanation and consequence through millions of templates.
3. **Refinement:** It is polished by **RLHF**, which imposes a human preference for logical coherence and narrative flow, effectively pruning non-causal outputs.
4. **Nature:** This understanding is **linguistic and schematic**. It often mirrors the force dynamics and implicit biases of human language, but can remain brittle when faced with novel physical interactions or rigorous counterfactual logic.

Overall, these systems can simulate many of the linguistic patterns humans use when describing causes and effects. That makes them useful for drafting, paraphrase, and extraction, but it should not be treated as evidence of intervention-level causal knowledge.

9. Comparative Data Tables

Table 1: Evolution of Causal Tasks and Metrics (2015-2025)

Era	Primary Focus	Methodology	Dominant Datasets	Typical Metric	"Native" Capability
2015-2018	Relation Classification	SVM, RNN, Sieves	SemEval-2010 Task 8, EventStoryLine	F1 Score (~0.50-0.60)	None (Pattern Matching)
2019-2021	Span/Context Extraction	BERT, RoBERTa	Causal-TimeBank, BioCausal	F1 Score (~0.72)	Contextual Recognition
2022-2025	Generative Reasoning	GPT-4, Llama, Instruction Tuning	CausalTalk, CRASS, ExpliCa	Accuracy, Human Eval	Generative/Schematic

Table 2: Performance on Causal Benchmarks (Selected Studies)

Benchmark	Task Description	Model Class	Performance Note	Source
SemEval Task 8	Classify relation between nominals	BERT-based (BioBERT)	~0.72-0.80 F1 (High accuracy on explicit triggers)	
CRASS	Counterfactual "What if" reasoning	GPT-3.5 / Llama	Moderate baseline; significantly improved with LoRA/PEFT	
CausalProbe	Causal relations in *fresh* (unseen) text	GPT-4 / Claude	Significant drop compared to training data; suggests memorisation	
Implicit Causality	Predicting subject/object bias (John amazed Mary)	GPT-4	High alignment with human psycholinguistic baselines	
Force Dynamics	Translating "letting/hindering" verbs	GPT-4	High accuracy in preserving agonist/antagonist roles	

Table 3: Key Instruction Tuning Datasets Influencing Causal Capability

Dataset	Content Type	Causal Relevance	Mechanism of Training	Source
FLAN	NLP Tasks -> Instructions	High (COPA, e-SNLI templates)	Explicitly maps "Premise" -> "Cause/Effect" in mixed prompts	
OIG	Open Generalist Dialogues	High (Advice, How-to)	Teaches Means-End reasoning (Action -> Result)	
Dolly	Human-generated Q&A	High (Brainstorming, QA)	Reinforces human-like explanatory structures	
CausalTalk	Social Media Claims	High (Implicit assertions)	Captures "gist" causality in informal discourse	

Related

- chapter intro