CAUSAL MAP GARDEN

!!! just my notes

Contents

🌻 Automated causal mapping can help answer evaluation questions

Chapter contents.

23 Aug 2025

Question for Step 3 - can automated causal mapping help answer evaluation questions?: An overview map was produced which included over 40% of the causal claims identified within the transcripts, using just 11 relatively broad factor labels.

The most central factor with the highest number of citations was Economic stress, which is a plausible result, with plausible connections to other factors.

We can use the map to identify and weigh up the evidence for contributions from and to individual factors. For example, the major contributions to Economic stress are Government policy and Covid-19, as well as “self-loops” mentioned by 46 sources, i.e. where one aspect of Economic stress was seen as causing another.

All such results depend on the (not automated) decisions made during the clustering process: how many clusters to use, whether to intervene in labelling, etc. This situation is closely parallel to decisions facing a statistician who has to identify variables for, say, structural equation modelling (Goertz 2020).

Comparison of citation frequency across timepoints was able to show that some links were mentioned significantly more than others, illustrating how this kind of map could be used to explore changes in systems (or in mental models of systems) over time.

Based on the provided factor rows, here is the analysis for rewriting the labels.

Good Minimum Set of Labels These labels cover the vast majority of the rows by grouping specific roles into broader categories.

- **[Respondent]** * *Covers:* The PhD student, "I", "My", personal feelings, skills, health, and career plans.
- *Examples:* "Coping mechanisms", "Anxiety", "Skill acquisition", "Motivation", "Work-life balance".
- **[Supervisor]** * *Covers:* PI, Promoter, Advisor, Boss, Group Leader.
- *Examples:* "Advisor support", "PI lack of presence", "Promoter management style", "Supervisor behavior".
- **[Peers]** * *Covers:* Colleagues, Other students, Postdocs (when acting as colleagues), Lab members.
- *Examples:* "Colleague behavior", "Peer support", "Postdoc mentorship", "Social exclusion".
- **[Institute]** * *Covers:* VIB, University, Center, Department, Management, HR, Admin, Committees, Doctoral School.
- *Examples:* "Administrative support", "Institute culture", "VIB training programs", "HR support".
- **[Lab]** * *Covers:* The immediate research group, team dynamics, and physical lab environment.
- *Examples:* "Lab atmosphere", "Group dynamics", "Toxic lab environment", "Resource allocation inequality".
- **[Research]** * *Covers:* The project, experiments, data, and scientific reality (used when a human actor is not the primary driver).
- *Examples:* "Experimental failures", "High data volume", "Project uncertainty", "Scientific discovery".
- **[Career]** * *Covers:* The job market, industry vs. academia, and future prospects (abstract).
- *Examples:* "Academic job market decline", "Industry sector characteristics", "Career prospects".

Other Actors to Consider These are distinct enough in the data that you might want separate labels for them, though they could be collapsed into the minimum set if necessary.

- **[Mentor]** * *Reasoning:* The data frequently distinguishes between a formal Supervisor/PI and a "Mentor" (who might be a postdoc or external).
- *Examples:* "Absence of VIB mentor", "Good mentor behavior", "Mentorship and role models".
- **[External]** * *Reasoning:* Actors outside the professional sphere or the institute.
- *Examples:* "Family", "Partner", "External collaborator", "External service provider".
- **[Support Staff]** * *Reasoning:* Distinct from "Institute" management and "Peers" doing research.
- *Examples:* "IT staff", "Technicians", "Core facility staff".

Labels That Do Not Fit the Schema The following labels describe abstract concepts, broad contexts, or interview artifacts that do not easily accept a specific Actor or Object label without losing meaning or forcing a fit.

Contextual/Abstract Factors: * "Passage of time" * "External crisis" (referring to COVID-19) * "Economic context" * "Societal gender norms" * "Nature of science" * "Demographic characteristics" (e.g., "being a guy vs girl") * "Cultural differences" (Abstract societal concept) **Interview Artifacts (AI-**

related): * "AI interviewer role constraints" * "AI refusal to answer off-topic query" * "Interviewer demographic inquiry" * "Interviewer performance" * "Broad interviewer questioning"

References

Goertz (2020). *Social Science Concepts and Measurement*. Princeton University Press.

PAGES IN THIS CHAPTER

🌻 Automated causal mapping can successfully code causal information

23 Aug 2025

Question for Step 2 - can automated causal mapping successfully code causal information?: Automated coding was able to identify causal claims made by respondents. The coding was noisy, with 35% dropping at least one quality point, but with no evidence of *systematic* errors. This level of precision is adequate for sketching out “causal landscapes” but would not be for high-stakes evaluations without additional manual correction. The accuracy can also be substantially improved by getting the AI to revise its work, (see redacted). This procedure still involves the researchers making significant high-level decisions in the formulation of the coding instructions as well as, before analysis, in clustering similar factor labels into groups. We believe this coding approach using genAI represents a significant improvement over the more hard-coded approaches for identifying causal relationships expressed in text (Dunietz 2018)Yang et al., 2022), and provides a more detailed, section-by-section coding which relies less on using AI as a black box to identify themes for initial coding (Jalali & Akhavan 2024) or to identify a global map (Graham 2023).

References

Dunietz (2018). *Annotating and Automatically Tagging Constructions of Causal Language.*

Graham (2023). *Using ChatGPT for Foresight: Futures Wheel.* https://medium.com/@christian.graham_49279/using-chatgpt-for-foresight-futures-wheel-8e79eecfe86b.

Jalali, & Akhavan (2024). *Integrating AI Language Models in Qualitative Research: Replicating Interview Data Analysis with ChatGPT*. https://doi.org/10.1002/sdr.1772.

Existentialists

21 Aug 2025

flashcards

Sartre shared many core ideas with other existentialists, but there were notable differences:

- **Simone de Beauvoir**: :: She expanded existentialism into feminist theory, arguing that women's oppression is a form of "bad faith." She emphasized that women must embrace their freedom and agency to overcome socially constructed roles.
- **Albert Camus**: existentialist? ? Though often associated with existentialism, Camus rejected the label. He focused on the concept of the "absurd," the conflict between humans' search for meaning and the indifferent universe. While Sartre emphasized freedom and choice, Camus highlighted the need to embrace the absurd and live with it.
- **Martin Heidegger**: ? An important influence on Sartre, Heidegger focused on "Being" rather than individual freedom. His concept of "thrownness" (Geworfenheit) described how individuals are always situated in a world not of their own making. While Sartre emphasized radical freedom, Heidegger focused on the interplay between individual existence and the world.
- **Karl Jaspers**: ?
- Jaspers emphasized the importance of "limit situations"—moments of existential crisis that force individuals to confront the fundamental questions of existence. While Sartre focused on freedom and choice, Jaspers leaned more towards the transcendental aspects of human experience.

"Existence precedes :: essence" is essentially saying that we aren't born with a predetermined purpose or nature. Instead, we create our own meaning and values through our choices and actions. The philosophical language can be dense and off-putting, but the core idea is straightforward: we define who we are by what we do, not by any inherent essence or destiny.

🌻 JMDE AI and transformative evaluation in the polycrisis

21 Aug 2025

Evaluation takes effort. It has a cost. Generative AI promises to make evaluation radically cheaper and easier to implement. Is this a good thing or a bad thing? In practice, of course, there is no simple answer. Evaluation takes place in many different contexts. It can be *adaptive*, helping people and organisations learn how to develop and optimise public and private activity such as the delivery of services; making evaluative judgements about what (not) to change, how, when, why and for whom. It can involve not only an assessment of facts but also a consideration of values (which? whose?). It can even be *transformative* when xxx. But it can also be *maladaptive* because it uses invalid information or makes invalid judgements or does not reflect the values of stakeholders. And then, there are all the *potential* use cases, where evaluation might take place but does not. These four possibilities are reflected in the rows of the table below. The cells contain some examples:

- A) every day, a teacher informally notices her students' mood and adapts her lesson accordingly. She has been doing it for years and hardly notices she is doing it.
- B1) A bike rental firm asks users to click a smiley or a frowny icon to say how happy they were, but it does nothing with this information.
- B2) As above, but the firm monitors changes in the proportion of frowny icons to identify locations or services which may need intervention. But it uses this to penalise regional staff without a causal understanding of underlying factors.
- C) a city council commissions an evaluation of its recycling programme from an external evaluator.
- D) A VOPE consults its members on how to make national evaluation practice more transformative

	informal	formal / internal	formal / external
transformative		D	
adaptive	A		C
maladaptive		B2	
potential		B1	

What might "introducing AI" mean in these cases? Anything could happen.

In case C the council might decide to use AI both to design and implement a new, realtime system which makes the external evaluator redundant. They might do this by transferring responsibility to an internal

evaluator to manage the process and vouch for its continuing validity. Or a manager might decide to do without any professional input, persuaded by the confident-sounding conclusions coming out of the new AI system. Trust is key. In order to to vouch for the robustness of the procedure, what do we rely on?

- an evaluation professional ?
- generic outside advice, e.g. from a management consultancy?
- a software provider (Microsoft? Google?)
- another AI tasked with validating the first?
- nothing at all: we simply rely on the confident-sounding outputs of the AI?

In case B the firm might redesign its system to use AI to make better, automated, realtime judgements about issues, causes and remedies in a way which is transparent and leaves room for employees and staff to respond, discuss and negotiate. Here we can dimly see a possible new role for evaluation professionals, to help co-design (and vouch for) new automated or semi-automated evaluation systems.

In case A it is not hard to imagine an AI system which monitors the students' social media feeds, tone of voice, eye movements as well as performance on set tasks, and gives the teacher realtime suggestions for pacing and adapting the class overall and for individuals. But most parents, students and teachers would (at the moment) likely be horrified at such a suggestion and it is doubtful if could be adaptive.

The bottom line in our table contains the largest potential (for adaptation, for transformation but also for poor or counterproductive solutions). At the extreme, a manager might engage an AI agent to sift through an entire organisation's workflows, identify areas which are currently not evaluated at all, and for each one, engage AI agents to suggest interoperable, cheap and robust solutions, and then connect them to tools to actually implement this new comprehensive suite of evaluation services. Of course, the implementation on paper might be laughably cheap, the real-life implementation is all about the human interface and might be arbitraily disruptve and expensive and nmaladaptive.

The day is probably not far off when a manager can open their computer one orning to be greeted with a essage like (n.d.) morning, i have created a system-level valuation suite while you were sleeping. Would you like to switch it on or do you want to bother with reviewing it first? Assuming you would likek to switch it on, will you wat it to work completely independetly or do you want to be infored about top-evel dcisions like hirig and firing?

🌻 Marina et al Rethinking rigour to embrace complexity in peacebuilding evaluation

22 Apr 2025

! bricolage

The article critiques dominant peacebuilding evaluation models for their rigid conceptions of rigour and calls for more adaptive, participatory approaches suitable for complex contexts. It introduces an **inclusive rigour framework** based on three interrelated domains:

1. **Methodological bricolage** – Flexible, pragmatic mixing of methods tailored to context and evolving questions.
2. **Meaningful participation and inclusion** – Emphasizing power dynamics and involving marginalized voices in co-creating evaluation processes.
3. **Utilisation and impact** – Ensuring evaluations produce actionable, contextually grounded learning for diverse stakeholders.

These are supported by enabling environments (e.g. equitable partnerships, supportive institutions) and evaluator competencies (e.g. reflexivity, political awareness).

Three case studies (Mali, Colombia Co-Inspira, Colombia Everyday Justice) illustrate application of the framework and highlight tensions between methodological rigor, participatory inclusion, and institutional expectations. The authors advocate redefining rigour as iterative, responsive, and grounded in local realities.

> “Rigour here is not defined by methodological choice alone, but rather, relies on an active view of evolving methodological choices throughout an iterative process” (p. 2).
> “Inclusive rigour… centres plural perspectives and use value over rigid hierarchies of evidence” (p. 5).

But in the slides it says

**

The values we ground the design in are:

- Seeking equity, and this was heavily covered in the previous training on rigor, recognizing participatory processes will increase the rigor of causal findings, but we will also explore it in this training as well.
- Informing strategy, and this was also covered in the previous training with a focus on the types of learning spaces we need to create. We'll lightly touch on it here.
- Embracing complexity, and we will discuss how to do this through our bricolage (or combining) of methods. This is very much a focus of today's training.

**

NORELLIEXPLANATORYLEARNINGEMPIRICISM

18 May 2025

Extracted Annotations (20/04/2022, 07:14:49)

"We formulate the challenge of creating a machine that masters a language as the problem of learning an interpreter from a collection of examples in the form (explanation; observations). The only assumption we make is this dual structure of data; explanations are free strings, and are not required to fit any formal grammar. This results in the Explanatory Learning (EL) framework described in Sec. 2" (Norelli et al 2022:2)

"Critical Rationalist Networks (CRNs), a family of models designed according to the epistemological philosophy pushed forward by Popper (1935). Although a CRN is implemented using two neural networks, the working hypothesis of such a model does not coincide with the adjustable network parameters, but rather with a language proposition that can only be accepted or refused in toto. We will present" (Norelli et al 2022:2)

"problem, where finite automata take the role of explanations, while regular sets are the phenomena. More recently, CLEVR (Johnson et al., 2017) posed a communication problem in a universe of images of simple solids, where explanations are textual and read like "There is a sphere with the same size as the metal cube". Another recent example is CLIP (Radford et al., 2021), where 400,000,000 captioned internet images are arranged in a communication problem to train an interpreter, thereby elevating captions to the status of explanations rather than treating them as simple labels3 . With EL, we aim to offer a unified perspective on these works, making explicit the core problem of learning an interpreter purely from observations." (Norelli et al 2022:3)

"many, the concept of explanation may sound close to the concept of program; similarly, the scientist problem may seem a rephrasing of the fundamental problem of Inductive Logic Programming (ILP) (Shapiro, 1981) or Program Synthesis (PS) (Balog et al., 2017). This is not the case. ILP has the analogous goal of producing a hypothesis from positive/negative examples accompanied by" (Norelli et al 2022:3)

"Such a model would assume that all the information needed to solve the task is embedded in the data, ignoring the explanations; we may call it a "radical empiricist" approach (Pearl, 2021). A variant that includes the explanations in the pipeline can be done by adding a textual head to the network. This way, we expect performance to improve because predicting the explanation string can aid the classification task. As we show in the experiments, the latter approach (called "conscious empiricist") indeed improves upon the former; yet, it treats the explanations as mere data, nothing more than mute strings to match, in a Chinese room fashion (Searle, 1980; Bender & Koller, 2020)" (Norelli et al 2022:6)

Results

22 Aug 2025

The sub-headings within each question form our criteria for answering that question.

Question 1: can an AI interviewer gather causal information at scale?

Efficiency

As we were still experimenting with the process, it took us around 8 hours to write, test, deploy and monitor the interviews.

We spent around $40 on API fees, including both tests and real interviews. The time and cost involved were significantly less than what it would have taken for humans to create an interview guideline and interview the same number of participants.

Validity

This is a difficult question to answer fully. However, in the interview prompt, we instruct the AI to summarise the conversation at the end of the interview and ask the respondent to verify its accuracy. We can use these answers to make a rough assessment of how valid the original summaries were: if the interviewee expresses no dissatisfaction, we can assume that the interviewer successfully elicited valid information.

The final section of all 163 interviews was analysed. We classified each interview into 3 groups:

1. No summary provided;
2. The respondent explicitly expressed dissatisfaction and/or asked for changes in the summary;
3. The respondent finished the interview and did not explicitly express dissatisfaction nor ask for changes in the summary.

78.5% of the respondents (**group 3**) didn't ask for changes in the summary, implying at least no dissatisfaction with what the AI produced (128 out of 163 interviews). Only in 7 interviews (4.29%) did the interviewee ask the AI to change or correct something in the summary, and/or the respondent explicitly expressed dissatisfaction (**group 2**). Of these, three then explicitly expressed satisfaction with the revised summary offered by the AI. The other 25 interviews (15.3%) were not summarised (**group 1**), mainly due to the participants breaking off before the end of the interview.

We used a much simpler architecture to manage the interview process than Chopra and Haaland (2023), however, our interviews were much shorter than theirs (their average interview length was about 30 minutes), raising the question of whether longer interviews might need more elaborate management architecture.

Question 2: Can automated causal mapping code causal information at scale?

Efficiency

It took around 5 hours to write and test the coding instructions and validate the results.

The cost of using the API was around $20.

Recall

Recall can be defined as the extent to which the AI finds "all" the causal links (Resnik & Lin 2010).

We made a separate assessment of the number of links "really" present within each interview, a "ground truth" of 1154 links. In comparison, the automated coding identified 1024 links, or 89%. However this is before assessing which of those codings were correct: the precision of the links, as follows.

Precision

Precision can be defined as the proportion of the identified links which were accurate/correct (Resnik & Lin 2010). To define "correct" we used the following informal criteria, which were assessed for each link by the second author:

1. The cause and effect in each link correctly name phenomena which are named in the text;
2. The coding represents an actual causal claim within the text (rather than, for example, merely events listed in sequence);
3. The coding represents a factual claim rather than a wish or hypothetical statement.
4. The coding is in the correct direction (cause to effect).

We gave each causal link a 0-2 score on the four criteria of precision as detailed in the Supplementary Material. 65% of the links had a perfect score, and 72% dropped only one point (a "not sure" on only one criterion). The errors we identified seem to take place approximately at random, except that there were more errors with causal claims which human analysts themselves judged to be difficult to code.

A more systematic assessment of the coding process on a real-life dataset had similar results and is currently in press (redacted).

Question 3: can automated causal mapping help answer evaluation questions?

Can an overall causal map be generated which includes much of the information?

The map in Figure 1 is filtered to show only the top 11 factors (in terms of the number of respondents mentioning them); links mentioned by only one source are also removed, meaning many less frequently mentioned factors and links are not shown.

We introduce a measure which we call **coding coverage**: given any map based on any recoding or filtering of the original data, what percentage of the original codings are included? There are balances to be struck: a map with more factors will usually have higher coverage but will be harder to understand and less useful. More homogeneity in sample and theme usually mean higher coverage. Very granular clustering will mean lower coverage or a larger map.

The first result can be seen in Figure 1. This map contains only 11 factors but covers 42% of the raw causal claims.

Insert Figure 1 around here.

Most (113 out of 136) sources have contributed at least some citations to this summary map. The numbers on the factors and links (and the sizes of the factors and the widths of the links) represent the number of sources mentioning each. Factors with darker backgrounds have proportionately more incoming than outgoing links: they have greater “outcome-ness”.

At this coarse level of “granularity”, many of the factors are bundles of cause-effect stories, as shown by the “self-loops” such as the 10 sources that mentioned links between different environment/climate change issues.

In this map, it is mostly not possible to distinguish between constituent factors with different valence or sentiment. For example, “military strengthening” and “military weakening” are two codes which have been included under “International conflict”. Indeed they are not so far from one another in the overall space of embeddings, something which is quite hard to understand from a positivistic, Cartesian point of view but which is perhaps more familiar to those more used to thinking in terms of "themes" than in terms of “variables”.

Face validity

Does the overall causal map present a plausible picture of the most important factors and how they influence one another (in the opinion of respondents)? Yes, even in the absence of a particular research

focus, this causal map has a lot to tell us about the causal worlds of the respondents.

“It’s the economy, stupid”: economic stress is mentioned by the largest number of sources and is central to most of the narratives. Covid-19 appears as a pure driver of economic stress.

Ability to answer other evaluation questions

Regarding the differences between timepoints, there were significant differences for several of the links. For example, of the five sources that mentioned the link from Political conflict to International conflict overall, all of them were from the third time-point, which is unsurprising considering the situation in Israel/Palestine at that timepoint.

In this analogue study, we did not have any additional information e.g. about the sociodemographic characteristics of the respondents which would have enabled us to look at differences between subgroups.

In a more realistic evaluation context, it would be possible to further investigate narratives about the causes and effects of specific factors of interest.

References

Resnik, & Lin (2010). *Evaluation of NLP Systems*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444324044#page=291.

VIA NEGATIVA — APOPHATIC REASONING

7 Oct 2025

You are absolutely right to make that distinction. While Popper's falsification is a famous and specific application of this way of thinking in science, the broader principle—that knowledge is advanced more robustly by negation and subtraction than by positive assertion—is indeed much older and more general.

This principle is most broadly known as the **Via Negativa** (Latin for "the negative way"), or **Apophatic Reasoning**.

The Via Negativa: Knowledge by Subtraction

The *via negativa* is a mode of thought that defines or understands something by what it is **not**, rather than by what it **is**. The core insight is that positive descriptions are often limited, speculative, and hard to agree on, whereas negative descriptions—ruling things out—can be more certain and build a stronger foundation for consensus.

The classic analogy is that of a sculptor. To create a statue of a horse, the sculptor doesn't "add" horseness to a block of marble. Instead, they **chip away everything that is *not* a horse**. The masterpiece is revealed through subtraction. 🗿

By Whom, Most Famously?

This idea has appeared in many domains, but here are its most famous proponents, long before Popper:

1. **Socrates and Platonic Philosophy:** The Socratic method is a quintessential example of the *via negativa* in action. Socrates rarely offers a definitive answer for what concepts like "justice" or "piety" *are*. Instead, his dialogues consist of him taking someone's proposed definition and showing, through questioning, its contradictions and inadequacies. The result is often *aporia*—an awareness of what you *don't* know. The knowledge gained is the elimination of bad definitions.

2. **Apophatic Theology (Negative Theology):** This is where the term is most famous. Philosophers and theologians like **Pseudo-Dionysius the Areopagite** (5th/6th century) and **Maimonides** (12th century) argued that any positive attribute we give to God (e.g., "God is good," "God is wise") is an inadequate human projection. We can't grasp the divine essence. Therefore, we can only speak truthfully about God by stating what God is *not*: not mortal, not finite, not corporeal, not ignorant. This clears the mind of false idols and leads to a more profound, if less defined, understanding.

3. **Nassim Nicholas Taleb (Contemporary):** The contemporary author who has done the most to popularize the *via negativa* as a practical principle for life and decision-making. In his books like *Antifragile*, he argues that we know much more about what is definitively harmful (what to subtract) than what is definitively beneficial (what to add). For example:
    - **Health:** It's easier to improve health by removing negatives (eliminating junk food, stopping smoking) than by finding the perfect "superfood."
    - **Wealth:** The surest path is avoiding ruin (not taking on catastrophic debt, not gambling foolishly).
    - **Knowledge:** We become wiser by identifying and eliminating our biases and false beliefs.

In your example of textual analysis, the *via negativa* means we create a space for rigorous and plausible interpretations by first chipping away and agreeing on the bad ones: those that are anachronistic, logically flawed, or unsupported by the text. What remains is a refined set of better, more defensible interpretations.