Our paper on an inductive workflow to gather and analyse evidence at scale.

22 Aug 2025

In our recent paper (Powell et al. 2025) we demonstrated that it is possible to gather evidence at scale about program theory and contribution simultaneously. Here is a preprint, and below is a summary.

Abstract

This article presents an artificial intelligence-assisted causal mapping pipeline for gathering and analysing stakeholder perspectives at scale. Evidence relevant to constructing a programme theory, as well as evidence for the causal influences flowing through it, are both collected at the same time, without the evaluator needing to possess a prior theory. The method uses an artificial intelligence interviewer to conduct interviews, automated coding to identify causal claims in the transcripts, and causal mapping to synthesise and visualise results. The authors tested this approach by interviewing participants about problems facing the United States. Results indicate that the method can efficiently collect and process qualitative data, producing useful causal maps that capture respondents' views as they evolve across time points. The article discusses the potential of this approach for evaluation. It also notes limitations and ethical concerns, emphasising the need for human oversight and verification.

Method

Our method comprised the following steps (following Tasks 1-3 according to Powell, Copestake, et al. (2023, p. 108-112):

Step 1: Conducting the chat interviews

This was a proof-of-concept analogue study. We employed online workers as respondents, recruited via Amazon's MTurk platform[3] (Shank, 2016). We decided to investigate respondents' ideas about problems facing the United States, as this generic theme was likely to elicit opinions from randomly chosen participants. This unsophisticated way of recruiting respondents means that the results cannot be generalised to a wider population in this case.

We had no specific evaluative questions in mind. We aimed to demonstrate a method which can be easily adapted to a specific research question.

A short semi-structured interview guideline was designed on the theme of 'What are the important current problems facing the USA and what are the (immediate and underlying) reasons for those problems?'. We aimed to construct an overall collective 'ToC' around problems in the United States. As it does not encompass a specific intervention this theory is not an example of a programme theory.

This interview guideline was implemented via an online interview 'AI interviewer' called 'Qualia',[4] which uses the OpenAI application programming interface (API) to control the AI's behaviour. Qualia is designed to elicit stories from multiple individual respondents, in an AI-driven chat format. Individual respondents are sent a link to an interview on a specific topic and, after consenting, are greeted by the interviewer. Rather than following a set list of questions, the interviewer is instructed to adapt its responses and follow-up questions depending on the respondents' answers, circling back to link responses and asking for more information as appropriate, focusing on the interview's objective mentioned above. These behaviours are based on the instructions written by the authors.

The respondents, who had the level of 'Master'[5] on Amazon's MTurk service, each completed an interview. The Amazon workers were given up to 19 minutes to complete the interview.

We repeated this interview at three different time points in September, October and November 2023, inviting approximately *N* = 50[6] respondents each time. The data from the three time points were pooled.

- The Research Question for Step 1 is: can an automated interview bot successfully gather causal information at scale?

Step 2: Coding the interviews

Step 2a: Constructing a guideline

Once the interviews were completed, we wrote instructions to guide the qualitative causal coding of the transcripts, in a radical zero-shot style: without giving a codebook or any examples. The assistant was told not to give a summary or overview but to list *each and every causal link or chain* of causal links and to ignore hypothetical connections (e.g. 'if we had X we would get Z'). We told the AI to produce codes or labels following this template: 'general concept; specific concept'. We gave no examples, but expected the AI to produce labels like: 'economic stress; no money to pay bills'. We call the combination of both parts a (factor) label.

The assistant was told also to provide a corresponding verbatim quote for each causal chain, to ensure that every claim could be verified. Codings without a quote which matched the original text were subsequently rejected, thus reducing the potential for 'hallucination'.

Step 2b: Coding

The final instructions were human-readable and could have been given to a human assistant. Instead, we gave these instructions to the online app 'Causal Map', which used the GPT-4 OpenAI API. As the transcripts were quite long (each around a page of A4 in length), each was submitted separately. The 'temperature' (the amount of 'creativity') was set to zero to improve reproducibility. The Causal Map app managed the housekeeping of keeping track of combining the instructions with the transcripts, watching out for any failed requests and repeating them, saving the causal links identified by the AI, and so on.

Step 2c: Clustering

The coding procedure resulted in many different labels for the causes and effects, many of which overlap in meaning. Even the general concepts (e.g. 'economic stress') were quite varied. The procedure for clustering these labels (including both the general and specific parts of the label) into common groups with their labels was a three-step process based on assigning to each of the original labels an embedding. An embedding is a numerical encoding of the meaning of each label (Chen et al., 2023) in the form of a vector (often visualised as a point in a high-dimensional space). For any two embedding vectors, cosine similarity can be calculated (measuring the angle between them) to quantify the semantic similarity between the labels they encode:[7]

1. *Inductive clustering*. First, we grouped the labels into clusters of similar labels using the hclust() function from the stats package of base (R Core Team, 2015).
2. *Labelling*. We then asked an AI to find distinct labels for each cluster. We also manually inspected these labels with regard to the original labels within each cluster and adjusted some of them.
3. *Deductive clustering*. We then discarded the original clustering, created embeddings for the new labels, and formed a new set of clusters, one for each of the new labels, assigning each original label to one of the new labels, the one to which it was most similar, providing the similarity was at least higher than a given threshold. This additional deductive step ensures that each member of each new cluster is sufficiently close in meaning to the new cluster label, rather than just to the other members of the cluster.

After each sub-step, we checked the AI's results to ensure that the instructions were being followed correctly and, if they were not, the instructions were tweaked or rewritten and tested again to ensure quality and consistency.

- The Research Question for Step 2 is: can automated causal mapping successfully code causal information at scale?

Step 3: Making useful syntheses of causal mapping data to answer evaluation questions

Standard filters (details on request) can be applied to the resulting data set of causal claims to create overview causal maps as a qualitative summary of the respondents' 'causal landscapes'. The primary aim is to construct a simple map with a not-overwhelming number of links and factors which captures a large percentage of the information given by the respondents. In addition, network metrics like centrality can be used to identify the factors which are most central within the network. To weigh up the evidence for the contributions made to a specific factor, we can list the evidence (the specific quotes from specific respondents) for direct and indirect links leading to it.

- The Research Question for Step 3 is: can automated causal mapping help answer evaluation questions?

Results and discussion

- *Question for Step 1. Can an AI interviewer successfully gather causal information at scale?*: Our AI interviewer was able to conduct multiple interviews with no researcher intervention at a low cost, reproducing the results of (Chopra & Haaland 2023). The interview transcripts read quite naturally and the process seems to have been acceptable to the interviewees.
- *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 per cent 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 Powell et al., forthcoming). 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; Dunietz et al., 2017; Jiang et al., 2023; Hooper et al., 2023; 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 and Akhavan, 2024) or to identify a global map (Graham, 2023).
- *Question for Step 3. Can automated causal mapping help answer evaluation questions?*: An overview map was produced which included over 40 per cent 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, that is, 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, and so on. This situation is closely parallel to decisions facing a statistician who has to identify variables for, say, structural equation modelling (Goertz, 2020: 136 ff).

Comparison of citation frequency across time points 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.

Caveats

Ethics, bias and validity

This kind of AI processing is not suitable for dealing with sensitive data because information from the interviews passes to OpenAI's (2024) servers, even though it is no longer used for training models.

Head et al. (2023) and Reid (2023) raise concerns about bias and the importance of equity in AI applications for evaluation, which have led to questions about the validity of AI-generated findings (Azzam, 2023). The way the AI sees the world, the salient features it identifies, the words it uses to identify them, and its understanding of causation are certainly wrapped up in a hegemonic worldview (Bender et al., 2021). Those groups most likely to be disadvantaged by this worldview are approximately the same who have least say in how these technologies are developed and employed.

AI is developing quickly: new models and techniques become available every month. However, we believe that any tools which genuinely add to knowledge should use procedures which are broken down into workflows consisting of simple individual steps, so that, humans can understand and check what is happening.

Interviewing

Researchers should carefully consider whether the interview subject matter is compatible with this kind of approach. For example, the AI may miss subtle cues or struggle to provide appropriate support to respondents expressing distress (Chopra and Haaland, 2023; Ray, 2023). We recommend that interview guidelines are tested and refined by human interviewers before being automated. No automated interview can substitute for the contextual information which a human evaluator can gain by talking directly to a respondent, ideally face-to-face and in a relevant context.

There is likely to be a differential response rate in this kind of interview: some people are less likely to respond to an AI-driven interview than others, and this propensity may not be random.

Causal mapping

Causal mapping is not at all suited for estimating the strength of causal effects: it can reveal the *strength of the evidence* for the influence of X on Y but this is not to be confused with the *strength of the effect* itself. There can be strong evidence for a weak link and vice versa.

Auto-coding

The work of the AI coder and clustering algorithms are not error-free. The coding of individual high-stakes causal links should be checked. In particular, there is a danger of accepting inaccurate results which look plausible.

This approach does not nurture substantive, large-scale theory-building of the kind expected, for example, in grounded theory (Glaser and Strauss, 1967). However, it can do smaller-scale theory-building in the sense of capturing theories implicit in individuals' responses.

This pipeline relieves researchers of much of the work involved in coding, but it is not fully autonomous. The human evaluator is responsible for applying the techniques in a trustworthy way and for drawing valid conclusions.

Potential

Qualitative approach

These procedures approach the stakeholder stories as far as possible without preconceived templates, to remain open to emerging and unexpected changes in respondents' causal landscapes.

Scalability and reach

The AI's ability to communicate in many languages presents an opportunity to reach more places and people, subject to Internet access and the AI's fluency in less common languages, and to include representative samples of populations.

The interview and coding processes are machine-driven and use zero temperature, so this approach should be mostly reproducible. Reproducibility opens the possibility of comparing results across groups, places and time points.

The low cost of coding large amounts of information means that it is much easier to develop, compare and discard hypotheses and coding approaches, something which qualitative researchers have previously been understandably reluctant to do.

Qualitative causality

These procedures have the potential to help evaluators answer evaluation questions which are often causal in nature, like: understanding stakeholders' mental models; judging whether 'their' ToC matches 'ours'; investigating 'how things work' for different subgroups of stakeholders; tracing impact from mentions of 'our' intervention to outcomes of interest; triaging the key outcomes in stakeholders' perspectives.

In summary, this kind of semi-automated pipeline opens up possibilities for monitoring, evaluation and social research which were unimaginable just 3 years ago and are well suited to today's challenging, complex problems like climate change and political and social polarisation. Previously, only quantitative research claimed to produce generalisable knowledge about social phenomena validly and at scale, by turning meaning into numbers. Now, perhaps, qualitative research will eclipse quantitative research by bypassing quantification and dealing with meaning directly, in somewhat generalisable ways.

Further work

We have tried to demonstrate a semi-automated workflow with which evaluators can capture stakeholders' emergent views of the *structure* of a problem or programme at the same time as capturing their beliefs about the *contributions* made to factors of interest by other factors. We have presented this approach via a proxy application but have since applied it in real-life research. Many challenges remain, from improving the behaviour of the automated interviewer through improving the accuracy of the causal coding process to dealing better with valence (e.g. distinguishing between 'employment', 'employment issues' and 'unemployment'). Perhaps, most urgently needed are ways to better understand and counter how LLMs may reproduce hegemonic worldviews (Head et al., 2023; Reid, 2023).

Related

- chapter intro

References

Chopra, & Haaland (2023). *Conducting Qualitative Interviews with AI*. https://doi.org/10.2139/ssrn.4583756.

Powell, Cabral, & Mishan (2025). *A Workflow for Collecting and Understanding Stories at Scale, Supported by Artificial Intelligence*. SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/13563890251328640.