A workflow for collecting and understanding stories at scale, supported by artificial intelligence

Steve Powell

Causal Map Ltd., UK

Gabriele Caldas Cabral

Causal Map Ltd., UK

Hannah Mishan

Bath SDR, UK

Corresponding author:

Steve Powell, 49, Highdale Road, Clevedon BS21 7LR, UK. Email: steve@pogol.net.

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. The article discusses the potential of this approach for evaluation, enabling rapid, large-scale qualitative analysis. It also notes limitations and ethical concerns, emphasising the need for human oversight and verification.

Keywords

AI-assisted qualitative research, causal mapping, automated interview analysis, AI-interviewing, contribution analysis

Motivation and background

To evaluate a programme, the evaluator can use contribution analysis (CA) (Mayne, 2012, 2015). We start with a programme logic or theory of change (ToC), consisting

of possible pathways from interventions to outcomes, and collect existing or new evidence for each link. However, evaluators can often not assume that the ToC underpinning a programme aligns with the realities on the ground, or they may uncover outcomes not anticipated in the original programme design – see Koleros and Mayne (2019). We have argued (Powell et al., 2024: 114) for a generalisation of CA in which 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 (necessarily) having a prior theory. In this sense, following Mayne, 'program theory' need not be something that any person necessarily possessed or articulated at the time, but is something which can be approximated and improved during the evaluation process.

(Re-)constructing programme theory empirically in this way is an essentially open-ended, qualitative problem. Closed data collection methods are not suitable because we cannot measure what we do not yet know. Open-ended, qualitative methods to (re-)construct a theory are notoriously time-consuming and are usually heavily influenced by researcher positionality (Copestake et al., 2019).

Powell et al. (2024) present this task as gathering and synthesising evidence about 'what influenced what' (p. 108), evidence which is simultaneously about theory or structure and contribution. Each piece of evidence may be of differing quality and reliability and about different sections of a longer pathway, or multiple interlocking pathways, and may come from different sources who see and value different things.

Causal mapping as part of the solution

One way to meet the challenge of simultaneously assessing theory and contribution is by using causal mapping (Ackermann and Eden, 2004; Axelrod, 1976; Eden et al., 1992; Hodgkinson and Clarkson, n.d.; Laukkanen and Wang, 2016; Powell et al., 2024) – the collection, coding and visualisation of interconnected causal claims. Causal mapping can be seen as unifying visual and text-based approaches and involves identifying causal claims within a set of texts. Each piece of text which contains information that a source S claims that X causally influences Y is coded: represented as a directed link or arrow in a network, connecting node X to node Y. Along with each link, the information about the origin of this claim, the source S, is also noted. Causal mapping provides a structured approach of synthesising (some of) the meaning of large quantities of text, from interviews or documents, providing a relatively generic way to extract and uncover emerging meaning. We can see causal mapping as a form of qualitative data analysis (QDA) in which the fundamental act of coding requires creating not a single tag or theme but an ordered pair of tags, the cause and the effect.

This evidence may (a) be collected systematically from relatively homogeneous sources, as in QuIP (Copestake et al., 2019) or (b) be a mix, consisting, for example,

of some systematic interviews, some official data and some reports from a monitoring system. In this article, we focus only on Option A.

If we broaden our view beyond any specific results chain, causal mapping can also be useful to evaluators who need to know how groups of people see the world, what these views of the world have in common and what distinguishes them, perhaps to make informed recommendations for programme improvement (Copestake et al., 2019).

We sometimes use the phrase ‘causal landscapes’ to describe the object of causal maps, to emphasise that we are most interested in the contents and layout of the broader world in which stakeholders understand themselves to be living.

We believe that causal mapping should carefully distinguish between evidence for a causal link and the causal link itself. While causal mapping can help the evaluator to identify, code and synthesise the evidence for causal connections, the evaluative judgement about whether one thing causally influences another is left to the evaluator (Better Evaluation, 2024).

Constructing causal maps by gathering and coding individual interview transcripts (rather than by consensus in groups; Ackermann and Eden, 2004; Barbrook-Johnson and Penn, 2022) can be a challenge for evaluators due to the complexity of the task and the resource-intensive and researcher-dependent nature of human-led coding. The use of artificial intelligence (AI)-driven tools is rapidly increasing among evaluators (Bohni Nielsen et al., 2024). AI can facilitate tasks by, for example, enabling virtual platforms to engage hard-to-reach populations or augment human capabilities by processing large data sets, auto-coding qualitative inputs or optimising data analysis (Eloundou et al., 2023).

Using causal mapping to gather evidence about structure/theory and contribution simultaneously

Our suggestion comprises the following steps (following Tasks 1–3 according to Powell et al., 2024: 108–112):

1. Gathering data by interviewing stakeholders about key issues of mutual interest (e.g. outcomes) and asking what drives these issues, and how they are interrelated with the drivers. For example, we can ask about outcomes and causes of outcomes and causes of causes. (We use the term ‘causal’ here in the loosest sense: we make causal connections every day using ordinary language when we say that one thing contributes to or drives or influences another, or makes, or might make, something else happen.) For applications using Option B (above), this step will look somewhat different, but we will not cover Option B here.

2. Code causal claims; we can then use causal mapping rules to identify causal claims within transcripts of these interviews. Each claim is a link between one cause or 'influence' factor and one effect or 'consequence' factor.[1] This will result in many individual causal maps, one per source/stakeholder.
3. Synthesise the individual causal maps into a causal network, showing common and diverging views, and then query the network to answer evaluation questions.

Scaling the approach

Gathering and processing sufficient data for this kind of approach has been a time-consuming task. Recent advances in AI raise the question: can we use AI to *gather* and *process* this kind of information at scale?

Previous work on each step

We are not aware of any prior work on our entire workflow, that is, using AI interviewers to gather data, construct causal maps and synthesise these to help answer evaluation questions. Here, we will refer to prior work on the three individual steps. Geiecke and Jaravel (2024) provide an end-to-end workflow with a single, general large language model (LLM) interviewer similar to ours and a suggestion for automated analysis, but their analysis approach is a global one involving the identification of themes across all transcripts and then identifying whether each theme is present in each transcript.

Step 1: Using an AI interviewer to gather causal information at scale

Data gathering with AI-powered chatbots has become popular in various businesses and industries. Chatbots have been used to gather customer data, feedback and to provide automated responses to frequently asked questions (Yuen, 2022) and in evaluation for tasks like gathering feedback (Bruce et al., 2024; Nielsen, 2023) or conducting 'laddering' interviews to investigate respondents' values (Rietz and Maedche, 2022). Chopra and Haaland (2023) report a recent use case in social research, suggesting that chat automations can be a valid and effective way to gather open-ended information from respondents. However, their approach involves a sophisticated set-up with multiple bots which are hard-coded into a team to monitor and steer the progress of the interview. Our experience has found hard-coding can be effective but difficult to adjust for different kinds of interviews. In this study, we leave the control of the interview, as well as developing and delivering the responses and new questions, all to a single automated interviewer. Geiecke and Jaravel (2024) present an approach similar to our own, using a single LLM agent to conduct useful social research interviews.

The term 'chatbot' for this kind of broader use of genAI to conduct natural-seeming conversations seems no longer appropriate, and we prefer the term 'AI interviewer'.

In the world of machine learning, a clear distinction can be made between supervised and unsupervised approaches (Ziulu et al., 2024). Using genAI to conduct interviews and code texts blurs this boundary. In our case, we developed our semi-generic instructions for interviewing, giving the AI instructions on how to behave, and how to make follow-up questions based on the interview objectives. Once the data collection is done, we create a separate genAI prompt to code causal links as a trial-and-error process, monitoring the quality of the coding post hoc. We did not have an explicitly stated ground truth about exactly how the interview should look or which causal claims were 'really' present within each text passage or how their causes and effects should be labelled, as we believe neither of these questions have a definitive answer; rather, we monitored AI's responses coding post hoc, iterating the prompt over many cycles to improve its performance. 'Prompt engineering' (Ferretti, 2023) like this can be considered a kind of supervision because it steers the AI's responses in a desired way.

Once the prompt was finalised, the interview AI was left to conduct interviews without further supervision. This prompt can remain broadly the same across different studies. However, the response of the AI can be highly sensitive to small differences in the 'prompt' and other settings (Jang and Lukasiewicz, 2023). Small adjustments made for specific studies, such as adjusting the instructions to focus better on research objectives, remain a vital point of human intervention.

Step 2: Using automated causal mapping to code causal information at scale

Making sense of texts by assigning codes or topics to text sections (or even entire documents) can be called thematic analysis (Braun and Clarke, 2006; Braun et al., 2021) or QDA (Lacey and Luff, 2001). Approaches to automating this process have moved on from topic modelling (Cintron and Montrosse-Moorhead, 2021) based on counting and clustering the words in the texts (Blei et al., 2003) to sentiment analysis (Roy and Rambo-Hernandez, 2021) and procedures which use LLMs to capture the meaning of longer sections of text (Sia et al., 2020; Ziulu et al., 2024).

Our task is more specific: identifying not just general meanings but also specifically causal relationships. Coding texts by hand for causal mapping has until recently been time-consuming work for trained analysts. Thus, it has remained a relatively niche approach. The limited sample size means it is difficult to make generalisations or comparisons between subgroups or across time points, reducing its utility for programme monitoring.

Earlier language models (Devlin et al., 2019) and other machine-based techniques have been used to identify causal relationships expressed in text (Dunietz, 2018; Dunietz et al., 2017; Hooper et al., 2023; Jiang et al., 2023) – for an overview, see Yang et al. (2022). However, these were highly specialised procedures which required 'training' the models. The advent of LLMs and genAI makes this process

much easier because we can rely on the model's inherent understanding of causality and directly ask the model to 'identify causal claims' rather than having to define exactly what this means.

The task of identifying causal claims is simpler than the broader problem of identifying themes in general, as with thematic analysis, because the task 'identify all the causal claims in this text' is more specific than 'identify all the themes in this text'. The latter approach would require a pre-reading of the text to know what kind of themes we might want to look for, which would leave the AI with too much freedom to make judgments for us and would also be more likely to expose any underlying model biases.

Evaluators (Davies, 2023; Ferretti, 2023) have recently been demonstrating the possibilities of AI in evaluation, for example, with asking AIs for summaries or global syntheses of texts including interviews (see Mason and Montrosse-Moorhead, 2023; Wachinger et al., 2024). It is even possible to ask an AI to synthesise the main causal links within a text and produce a diagram directly (Graham, 2023). However, when doing this, evaluators have to take care not to leave fundamental evaluative decisions, such as 'within this text, what are the most important claims' to the LLM. Evaluators should be wary of transferring the responsibility for making evaluative judgements to an unknown third party, the 'black box' of the AI (Choudhary et al., 2022). One way to navigate this complexity is to systematically break down larger, weakly specified tasks into multiple, smaller, better-specified tasks and also to clearly distinguish where AI adds value and where human insight is needed.

In this sense, we prefer approaches which use the power of genAI more transparently, as a low-level coding assistant in the tradition of QDA or thematic analysis who follows detailed instructions, leaving the evaluator with the responsibility of making evaluative judgements. This involves establishing coding guidelines designed to extract causal information from the documents with little guidance.

Jalali and Akhavan (2024) use one such approach. They first instruct the AI to construct a codebook (a list of salient causal factors) based on the whole text and then use this codebook to identify and code links. While they report good results, we see this as still leaving too much freedom for the AI to make its own global 'judgement' based on processing the whole text about what are the salient factors and therefore also being more exposed to bias inherent in the AI's 'worldview'.

There are different ways to create factor labels for the influence and consequence factors making up of each causal link. They may be:

- Specified in advance deductively in the form of a codebook:
- – Based on pre-existing theory.

- – Based on a preliminary assessment of the text to be coded, as with Jalali and Akhavan.
- According to a codebook which is developed iteratively during coding (Laukkanen, 1994) and QuIP (Copestake et al., 2019).
- Created 'In vivo' in whatever form is most suited to each coding. This means that there will be many different labels for the causes and effects, many of which are likely to overlap in meaning. We call this approach to causal coding 'radical zero-shot':[2] no codebook or examples are provided, and there is no mechanism to ensure consistent labels across the data set. This coding procedure maximises granularity but means that to build synthesis maps, it will be necessary to subsequently, retrospectively, cluster labels into sets with similar meaning and provide appropriate labels for the clusters. This is the approach we prefer and use in this article.

As with creating the instruction ('prompt') for the AI interviewer, developing an instruction to identify causal links was a trial-and-error process, involving monitoring the quality of the coding post hoc, making adjustments post hoc as we analysed the quality of the coding and identified errors or gaps.

Step 3: Using automated causal mapping to help answer evaluation questions

The extensive causal mapping literature provides many examples of its use to answer evaluation questions (see Powell et al., 2024: 110), for example:

- *Getting an overview of respondents' 'causal landscape'*. This can be useful for orientation or for particular tasks like triaging masses of information to identify key outcomes and possible causal pathways when planning an outcome harvesting (Wilson-Grau and Britt, 2012) or process tracing (Befani and Stedman-Bryce, 2017) project.
- *Weighing up evidence about contribution.* in particular, tracing back and comparing the possibly multiple contributory causes of an important outcome or consequence (Goertz and Mahoney, 2006), or examining effects of causes.
- Reporting key metrics of the causal network, for example, to reveal which factors are most central in the whole network or to identify feedback loops.
- Asking whether the empirical ToC matches the plan (Powell et al., 2023b: 7).
- Making comparisons between groups or across time points.

We have integrated the research questions into the method steps below and set out more detailed criteria for evaluating the research questions within the 'Results' section.

Method: The 'AI-assisted causal mapping pipeline'

We present a procedure which harnesses the power of AI to facilitate and augment evaluation practice (Eloundou et al., 2023) in three ways: first to carry out large numbers of automated, qualitative, online interviews, second, to automatically code the transcripts and third, to present overview causal maps.

Step 1: Conducting the chat interviews

This article presents results from 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 point in a space, such that, two labels with similar meaning are close in this space. For any two such vectors, a measure cosine similarity can be calculated representing the approximate similarity in meaning between the labels which 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

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 US$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 three 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.

Of the respondents (**group 3**), 78.5 per cent did not ask for changes in the summary, implying at least no dissatisfaction with what the AI produced (128 of 163 interviews). Only in seven 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 US$20.

Recall

Recall can be defined as the extent to which the AI finds ‘all’ the causal links (Resnik and 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 per cent. 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 and 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, e.g. 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. Sixty-five per cent of the links had a perfect score, and 72 per cent 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 data set had similar results and is currently in press (Powell et al., forthcoming).

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.

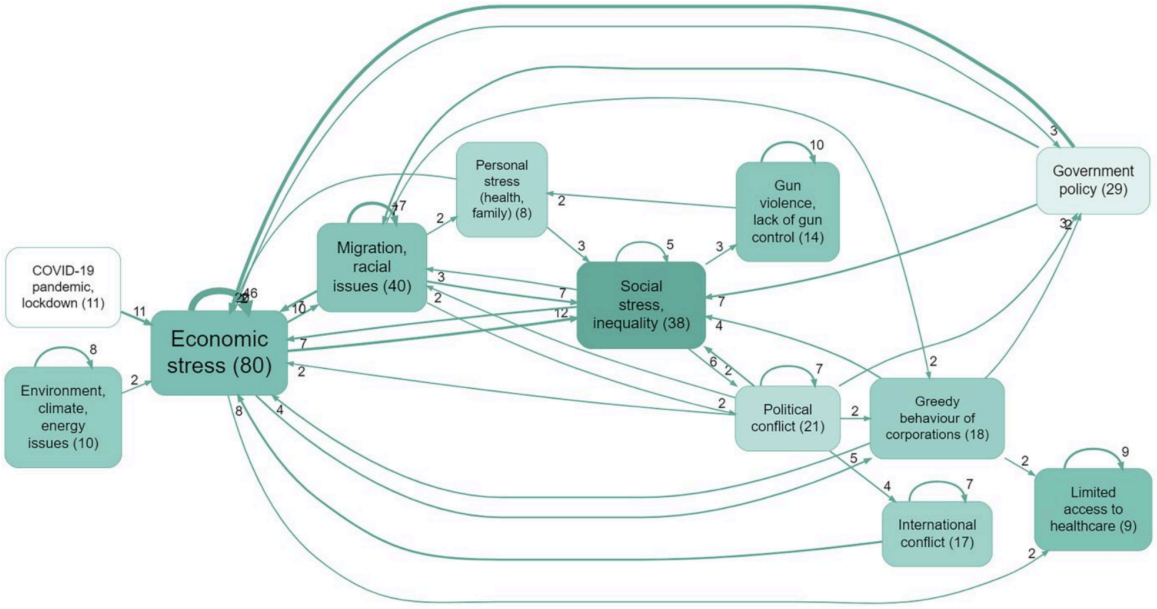

Figure 1. A high-level overview causal map.

Causal factors are automatically clustered as described in the Supplementary Material.

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 is 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 per cent of the raw causal claims.

Most (113 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’.[8]

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 valences or sentiments. 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 time points, 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 time point.

In this analogue study, we did not have any additional information, for example, 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.

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 and Haaland (2023) and Geiecke and Jaravel (2024). 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).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This study was independently financed by Causal Map Ltd.

ORCID iD

Steve Powell https://orcid.org/0000-0002-8776-9845

Gabriele Caldas Cabral https://orcid.org/0000-0003-0120-784X

Hannah Mishan https://orcid.org/0009-0003-8170-1430

Supplemental material

Supplemental material for this article is available online.

Notes

1.We prefer ‘influence’ and ‘consequence’ to ‘cause’ and ‘effect’ because they make it clear that we are not suggesting that we are ever likely to encounter a case in which one thing was the exclusive, deterministic cause of an effect, but rather than we are dealing with partial influences and partial contributions.

2.Coding tasks which provide a codebook or suggested themes for coding without providing any examples or training are often termed ‘zero-shot’.

3.A crowdsourcing marketplace that enables researchers and others to engage remote workers to carry out small tasks online.

4.Previously called StorySurvey.

5.Masters are ‘a specialized group of Workers who consistently demonstrate accuracy in performing a wide range of [tasks]’ (Amazon Inc, 2016).

6.A few respondents broke off the interview before actually completing. Their data were also included.

7.This procedure is described in more detail in the Supplementary Material.

8.The proportion of citations of outgoing links out of all the citations of a particular factor: a normalised version of the Copeland Score (Copeland, 1951). Factors with low outcomeness can be thought of as inputs or drivers.

References

Ackermann F and Eden C (2004) Using causal mapping: Individual and group; traditional and new. In: Pidd M (ed.) *Systems Modelling: Theory and Practice*. Chichester: Wiley, 127–45.

Amazon Inc (2016) Simplified masters qualifications. Available at: https://blog.mturk.com/simplified-masters-qualifications-137d77647d1c (accessed 4 December 2023).

Axelrod R (1976) The analysis of cognitive maps. In: Axelrod R (ed.) *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton, NJ: Princeton University Press, 55–77.

Azzam T (2023) Artificial intelligence and validity. *New Directions for Evaluation* 2023(178–179): 85–95.

Barbrook-Johnson P and Penn AS (2022) *Systems Mapping: How to Build and Use Causal Models of Systems*. Cham: Springer. Available at: https://link.springer.com/10.1007/978-3-031-01919-7 (accessed 14 May 2023).

Befani B and Stedman-Bryce G (2017) Process tracing and Bayesian updating for impact evaluation. *Evaluation* 23(1): 42–60.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, Virtual Event*, 3–10 March, 610–23. New York: ACM. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (accessed 14 November 2023).

Better Evaluation (2024) Causal mapping. Available at: https://www.betterevaluation.org/methods-approaches/methods/causal-mapping (accessed 3 August 2024).

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Bohni Nielsen S, Mazzeo FR and Petersson G (2024) Evaluation in the era of artificial intelligence. In: Rinaldi FM, Petersson GJ and Nielsen SB (eds) *Artificial Intelligence and Evaluation*. New York: Routledge, 1–12.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101.

Braun V, Clarke V, Boulton E, et al. (2021) The online survey as a qualitative research tool. *International Journal of Social Research Methodology* 24(6): 641–54.

Bruce K, Gandhi VJ and Vandelanotte J (2024) Emerging technology and evaluation in international development.
In: Rinaldi FM, Petersson GJ and Nielsen SB (eds) *Artificial Intelligence and Evaluation*. New York: Routledge: 21-23.

Chen S, Zhang H, Chen T, et al. (2023) Sub-sentence encoder: Contrastive learning of propositional semantic representations. *arXiv*. Available at: http://arxiv.org/abs/2311.04335 (accessed 14 November 2023).

Chopra F and Haaland I (2023) Conducting qualitative Interviews with AI*.: 72*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4583756

Choudhary S, Chatterjee N and Saha SK (2022) Interpretation of black box NLP models: A survey. *arXiv [preprint]*. DOI: 10.48550/arXiv.2203.17081.

Cintron D and Montrosse-Moorhead B (2021) Integrating big data into evaluation: R code for topic identification and modeling. *American Journal of Evaluation* 43: 109821402110316.

Copeland AH (1951) *A Reasonable Social Welfare Function* (mimeo, 1951). Ann Arbor, MI: University of Michigan.

Copestake J, Davies G and Remnant F (2019) Generating credible evidence of social impact using the Qualitative Impact Protocol (QuIP): The challenge of positionality in data coding and analysis. In: Clift BC, Gore J and Bekker S, et al. (eds) *Myths, Methods, and Messiness: Insights for Qualitative Research Analysis*. Bath: University of Bath, 17–29.

Davies R (2023) Evaluating thematic coding and text summarisation work done by artificial intelligence (LLM). *Rick on the Road*. Available at: http://mandenews.blogspot.com/2023/08/evaluating-thematic-coding-and-text.html (accessed 13 October 2023).

Devlin J, Chang M-W, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. Available at: http://arxiv.org/abs/1810.04805 (accessed 2 December 2023).

Dunietz J (2018) *Annotating and automatically tagging constructions of causal language*. PhD Thesis, Brandeis University, Waltham, MA.

Dunietz J, Levin L and Carbonell J (2017) The BECauSE Corpus 2.0: Annotating causality and overlapping relations. In: *Proceedings of the 11th linguistic annotation workshop*, Valencia, April, 95–104. Stroudsburg, PA: Association for Computational

Linguistics. Available at: http://aclweb.org/anthology/W17-0812 (accessed 14 November 2023).

Eden C, Ackermann F and Cropper S (1992) The analysis of cause maps. *Journal of Management Studies* 29(3): 309–24.

Eloundou T, Manning S, Mishkin P, et al. (2023) GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv*. Available at: http://arxiv.org/abs/2303.10130 (accessed 10 February 2025).

Ferretti S (2023) Hacking by the prompt: Innovative ways to utilize ChatGPT for evaluators. *New Directions for Evaluation* 2023(178–179): 73–84.

Geiecke F and Jaravel X (2024) *Conversations at scale: Robust AI-led Interviews with a simple open-source platform*. 4974382, SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=4974382 (accessed 25 November 2024).

Glaser BG and Strauss AL (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Berlin: Aldine de Gruyter.

Goertz G (2020) *Social Science Concepts and Measurement* (New and completely revised edition). Princeton, NJ: Princeton University Press.

Goertz G and Mahoney J (2006) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, NJ: Princeton University Press.

Graham C (2023) Using ChatGPT for foresight: Futures wheel. *Medium*. Available at: https://medium.com/@christian.graham_49279/using-chatgpt-for-foresight-futures-wheel-8e79eecfe86b (accessed 2 December 2023).

Head CB, Jasper P, McConnachie M, et al. (2023) Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation* 2023(178–179): 33–46.

Hodgkinson GP and Clarkson GP (n.d.) What have we learned from almost 30 years of research on causal mapping? Available at: https://www.irma-international.org/viewtitle/6745/?isxn=9781591403968

Hooper R, Goyal N, Blok K, et al. (2023) A semi-automated approach to policy-relevant evidence synthesis: Combining natural language processing, causal mapping, and graph analytics for public policy. Available at: https://www.researchsquare.com (accessed 14 November 2023).

Jalali MS and Akhavan A (2024) Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review* 40(3): e1772.

Jang ME and Lukasiewicz T (2023) Consistency analysis of ChatGPT. *arXiv [preprint]*. DOI: 10.48550/arXiv.2303.06273.

Jiang H, Ge L, Gao Y, et al. (2023) Large language model for causal decision making. *arXiv*. Available at: http://arxiv.org/abs/2312.17122 (accessed 16 January 2024).

Koleros A and Mayne J (2019) Using actor-based theories of change to conduct robust contribution analysis in complex settings. *Canadian Journal of Program Evaluation* 33(3): 52946.

Lacey A and Luff D (2001) Qualitative data analysis. *Trent focus Sheffield*. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=97dbcfff8631bff451fd4384345e19546f8b64dc (accessed 27 February 2024).

Laukkanen M (1994) Comparative cause mapping of organizational cognitions. *Organization Science* 5(3): 322–43.

Laukkanen M and Wang M (2016) *Comparative Causal Mapping: The CMAP3 Method*. New York: Routledge.

Mason S and Montrosse-Moorhead B (2023) Editors' notes. *New Directions for Evaluation* 2023(178–179): 7–10.

Mayne J (2012) Making causal claims. ILAC Brief 26. Available at: https://cgspace.cgiar.org/items/110c7654-2d4a-4f4e-9174-4f6d573b0e6f (accessed 20 March 2025).

Mayne J (2015) Useful theory of change models. *Canadian Journal of Program Evaluation* 30(2): 230.

Nielsen SB (2023) Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. *New Directions for Evaluation* 2023(178–179): 47–57.

OpenAI (2024) Enterprise privacy at OpenAI. *Enterprise Privacy at OpenAI*. Available at: https://openai.com/enterprise-privacy/

Powell, S., Copestake, J. and Remnant, F., 2024. Causal mapping for evaluators. Evaluation, 30(1), pp.100-119.

Powell S, Caldas-Cabral G and Remnant F (Forthcoming). AI-Assisted Causal Mapping: A Validation Study. *IJSRM*.

Powell S, Larquemin A, Copestake J, et al. (2023) Does our theory match your theory? Theories of change and causal maps in Ghana.
In: Simeone L, Drabble D and Morelli N, et al. (eds) *Strategic Thinking, Design and*

the Theory of Change: A Framework for Designing Impactful and Transformational Social Interventions. Cheltenham: Edward Elgar, 232–50.

Ray PP (2023) ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3: 121–54.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reid AM (2023) Vision for an equitable AI world: The role of evaluation and evaluators to incite change. *New Directions for Evaluation* 2023(178–179): 111–21.

Resnik P and Lin J (2010) Evaluation of NLP systems.
In: Clark A, Fox C and Lappin S (eds) *The Handbook of Computational Linguistics and Natural Language Processing*. Hoboken, NJ: Wiley, 271–95.

Rietz T and Maedche A (2022) Ladderbot – A conversational agent for human-like online laddering interviews. *SSRN Electronic Journal*. Epub ahead of print 21 March. DOI: 10.2139/ssrn.4062500.

Roy A and Rambo-Hernandez K (2021) There's so much to do and not enough time to do it! A case for sentiment analysis to derive meaning from open text using student reflections of engineering activities. *American Journal of Evaluation* 41: 1–17.

Shank DB (2016) Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *The American Sociologist* 47(1): 47–55.

Sia S, Dalmia A and Mielke SJ (2020) Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! *arXiv*. Available at: http://arxiv.org/abs/2004.14914 (accessed 26 February 2024).

Wachinger J, Bärnighausen K, Schäfer LN, et al. (2024) Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research*. Epub ahead of print 22 May. DOI: 10.1177/10497323241244669.

Wilson-Grau R and Britt H (2012) Outcome harvesting. *Cairo: Ford Foundation*. Available at: https://outcomeharvesting.net/wp-content/uploads/2016/07/Outcome-Harvesting-Brief-revised-Nov-2013.pdf (Accessed 20 March 2025.

Yang J, Han SC and Poon J (2022) A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems* 64(5): 1161–86.

Yuen M (2022) Chatbot market in 2022: Stats, trends, and companies in the growing AI chatbot industry. *Insider Intelligence*. Available at: https://www.insiderintelligence.com/insights/chatbot-market-stats-trends/

Ziulu V, Anuj H, Hagh A, et al. (2024) Extracting meaning from textual data for evaluation: Lessons from recent practice at the independent evaluation group of the World Bank. In: Rinaldi FM, Petersson GJ and Nielsen SB (eds) *Artificial Intelligence and Evaluation*. New York: Routledge, 287–308.

Steve Powell has been working in research and evaluation since 1995, with projects on a wide range of topics from psychosocial programming after disasters to social capital in the former Soviet Union. He holds a PhD in psychology, is fascinated by causal mapping and is co-founder of *proMENTE social research* in Sarajevo and *Causal Map Ltd* in Bristol.

Gabriele Caldas Cabral has an academic background in project management, international relations and international development with a research focus on South–South Cooperation. She is an Outreach Coordinator at *Causal Map Ltd*, where she concentrates on automating qualitative data collection and analysis for causal mapping.

Hannah Mishan has a background in international development, with a focus on sustainability. She is particularly interested in utilising causal mapping to improve the usability of evaluations. She works with both *Causal Map Ltd* and *Bath Social & Development Research*, and in this role, she enjoys using causal mapping to analyse and present research findings.