🌻 Qualia.

Chapter Contents.

> There are many different ways to collect data for causal mapping: Task 1 -- Introduction.

One of our favourite ways is with Qualia, our AI interviewer -- though Qualia can of course be used for other kinds of data collection, not just for causal mapping.

You might first want to look at the Qualia technical documentation. That documentation tells you what buttons to press and gives all the details of setting up, sharing and managing your interviews.

This chapter (which like the rest of this site is a constant work in progress) gives you the background:

- what research have we done on Qualia?
- how do you create a really great interview?

PAGES IN THIS CHAPTER

> expressing distress [@chopraConductingQualitativeInterviews2023; @rayChatGPTComprehensiveReview2023]. 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.

AI interviewing - the evaluator retains responsibility

> 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.

AI interviewing has potential - scalability, reach, reproducibility, causality

> **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.

AI interviewing needs 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 program 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 (for example distinguishing between "employment", "employment issues" and "unemployment"). Perhaps most urgently needed are ways to better understand and counter how LLMs may reproduce hegemonic worldviews [@reidVisionEquitableAI2023].

An AI interviewer can successfully gather causal information at scale

> **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 [@chopraConductingQualitativeInterviews2023; @anderssonTheoryChangeSustainable2024]. The interview transcripts read quite naturally and the process seems to have been acceptable to the interviewees.

CASA

> People are often more candid with machines than with other people. Why?

How Qualia copes with different languages

> There are two things to think about, the transcription service (necessary only if we enable the option for people to speak instead of type) and the AI interviewer service which provides interviewer responses.
>
> - Brazilian Portuguese should be fine for both.
> - Kurdish would require us using dedicated services for both, it probably wouldn't be worth it.
> - For Arabic variants (beyond Modern Standard), the situation is more tricky, but probably similar for both. As I understand the current state of affairs the problem the models have with Arabic variants is more about cultural adaptation rather than the language itself. For voice transcription we would probably need us to install a special model which would then reportedly be ok in Jordan, and for the chat interviewer service we'd probably use our standard gpt-4.1 as that is promising for Arabic variants. But we can't guarantee this would work.
> - Otherwise the top 50 or so languages in terms of how present they are on the internet should all work fine.
> - Although Qualia does a very good job of detecting / guessing the respondent's preferred language and adapting to that, we get best results if we don't do that but tell it in advance which language will be used -- but this means people who we expect to use, say, Portuguese are not then able to switch to, say, English.

It is possible to gather evidence at scale about program theory and contribution simultaneously - three steps

Our seamless stories workflow in practice

> Automating chat interviews with **Qualia**. Then using **Causal Map** to make sense of them. In-depth research was never this easy! A case study from Chile.

Qualia and data security

Qualia asks about USA problems, again

> How can we capture and visualise people’s mental models of a complex situation like the state of a nation? This week, as part of an EES webinar demonstrating our automated AI interviewer Qualia, we asked the participants to spend a few minutes being interviewed about problems facing the USA and the reasons for them, and the reasons for the reasons. Over 90 people did, with a mean of 13 messages per conversation. Details below.

Step 1 – Conducting the chat interviews

> 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" [@ferrettiHackingPromptInnovative2023] like this can be considered a kind of supervision because it steers the AI’s responses in a desired way.

Step 2a Coding the interviews – 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 (for example, “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.

Step 2b Coding the interviews – 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, etc.

Step 2c Coding the interviews – 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:

Using AI interviewing - beware of bias

[@headLargeLanguageModel2023] and [@reidVisionEquitableAI2023] 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 [@azzamArtificialIntelligenceValidity2023]. 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.

Your interview instructions have to be explicit

> Writing explicit interview instructions for our AI-interviewer Qualia (QualiaInterviews.com) is fascinating because you have to be explicit about everything, including how much you want the same questions asked every time and how much you want your AI assistant to chase topics down rabbit-holes.