CHAPTER

# Qualia

## CONTENTS

## Your interview instructions have to be explicit

> 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?

## The seamless workflow from AI interviews to causal map

An AI interviewer can successfully gather causal information at scale ▸

!

!

AI interviewing - beware of sensitive data ▸

Using AI interviewing - beware of bias ▸

AI interviewing - beware of suitability ▸

AI interviewing - the evaluator retains responsibility ▸

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

AI interviewing needs further work ▸

AI interviewing - beware of sensitive data

# AI interviewing - beware of sensitive data

## 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 servers, even though it is no longer used for training models ({OpenAI, 2024).

## References

{OpenAI (2024). *Announcing GPT-4o in the API! - Announcements*. https://community.openai.com/t/announcing-gpt-4o-in-the-api/744700.

# AI interviewing - beware of suitability

## 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 & 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.

## References

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

Ray (2023). *ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope.*. https://doi.org/10.1016/j.iotcps.2023.04.003.

# AI interviewing - the evaluator retains responsibility

### Autocoding

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 & 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.

## References

Glaser, & Strauss (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter.

# 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.

**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 timepoints.

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 three 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.

# AI interviewing needs further work

## 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 (Reid, 2023).

## References

Reid (2023). *Vision for an Equitable AI World: The Role of Evaluation and Evaluators to Incite Change*. https://doi.org/10.1002/ev.20559.

# 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 (Chopra & Haaland, 2023); (Andersson, 2024). The interview transcripts read quite naturally and the process seems to have been acceptable to the interviewees.

## References

Andersson (2024). *Theory of Change for Sustainable Business*. In *Theories of Change in Reality*. https://doi.org/10.4324/9781032669618-9.

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

# CASA

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

It is probably related to the "Computers Are Social Actors" (CASA) paradigm. This theory suggests that humans often interact with computers as if they were social beings. However, the perceived lack of genuine consciousness, feelings, and social judgment in AI can reduce the pressure to maintain a socially desirable persona. Candidates may feel that the AI is a less judgmental evaluator, leading to more straightforward and less embellished responses.

## 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.

- **The Language Capability Doubter**
  - A client who questions whether Qualia can effectively handle interviews in their target language or across multiple languages.
  - Qualia supports approximately the top 50 languages present on the internet, with particularly strong capabilities in major languages like Brazilian Portuguese.
  - For optimal results, we can configure Qualia to specifically operate in your target language rather than relying on automatic detection.
  - The system combines both transcription services (for spoken responses) and AI interviewer capabilities customized to your language needs.
  - Less common languages may require special considerations, we can evaluate feasibility for your specific language requirements.
  - Qualia's language capabilities allow for consistent interview quality across different markets, ensuring comparable data collection.
- **The AI Reliability Skeptic**
  - A client who is uncertain about the reliability, quality, and authenticity of AI-conducted interviews compared to traditional human methods.
  - Qualia operates on the best available new generative AI technology, producing consistent and friendly interviews that eliminate human interviewer variation.
  - The system can be precisely configured to follow your interview protocol, ensuring methodological rigor.

- We can provide demonstrations showing how Qualia handles different respondent types and interview scenarios.
- The AI interviewer can adapt to respondent answers while maintaining your research objectives, combining flexibility with consistency.
- Using Qualia allows you to conduct more interviews within your budget, significantly increasing sample size and explanatory power.

## 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.

At Causal Map we're thrilled because our seamless AI-supported workflow is finally coming together. Recently we helped colleagues at a University in Chile to complete a qualitative, explorative evaluation of the impact of a programme, using our automated interviewer **Qualia** to conduct the interviews and **Causal Map** to make sense of them.

This workflow means you can do **in-depth research** so much more **quickly** and **cheaply** than before while maintaining depth and quality, opening up new possibilities for understanding complex social issues.

## Background

DuocUC, a higher education institution in Chile, hired our consultancy to conduct QuIP-style interviews with Qualia and analyse them using the Causal Map app. The interviews were motivated by concerns about the gender gaps faced by women pursuing STEM careers at the university.

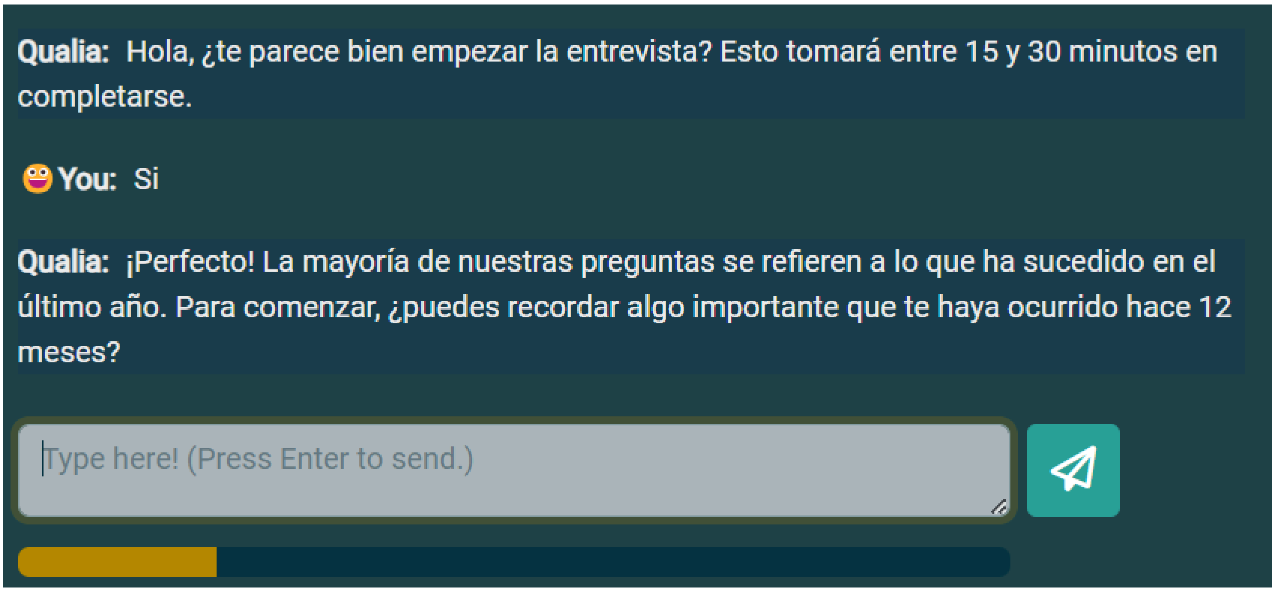
This study has been developed in the quality assurance department, as part of the institutional evaluation strategies, led by Felipe Rivera, Head of Academic Quality Evaluation.

We had a first meeting to understand what they wanted to find out, their research questions and the scope of the study and to determine the domains in which the interviews would be conducted.

After this, we started writing the instructions for Qualia to conduct the interviews, having a few iterations with the client’s team to come up with an interview structure that would suit them.

## Step 1: Setting up the interview in Qualia

- The instruction for the AI interviewer was similar to the instructions you could give to a human interviewer. And both the interview instructions and the interviews itself were conducted in Spanish.
- The AI asked questions about changes in 3 domains: educational experiences, professional development and relationship dynamics.
- We used GPT-4o which is the best AI model to date.



**Qualia:** Hola, ¿te parece bien empezar la entrevista? Esto tomará entre 15 y 30 minutos en completarse.

😜 **You:** Si

**Qualia:** ¡Perfecto! La mayoría de nuestras preguntas se refieren a lo que ha sucedido en el último año. Para comenzar, ¿puedes recordar algo importante que te haya ocurrido hace 12 meses?

## Step 2: Collecting stories with Qualia

- We sent the interview link to 50 people and were able to collect 32 interviews.
- We created special individual links to be able to track the interviews:
  - At Qualia, we don't store personally identifying information at all. But we can add a personalised key like &key=0003 to the end of the URL for each individual invitation.
  - And this allowed the researchers to keep track of who they sent which invitation to, so that they knew that e.g. key 0003 belongs to Claudia.
- We downloaded the interview results from Qualia and uploaded them into Causal Map.

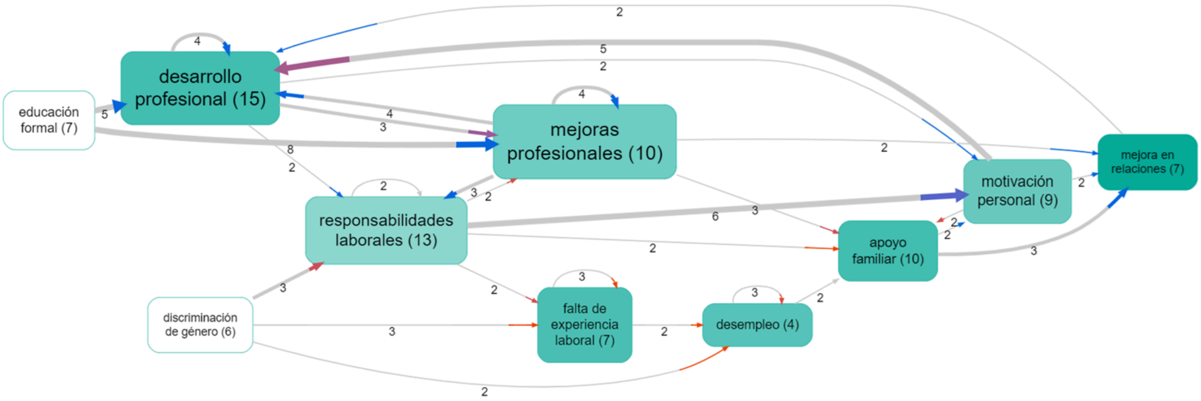
## Step 3: Analysing stories with Causal Map

- We used AI (GPT-4o) to identify each and every causal link in the interviews, and for each link, to label the cause and effect.
- We used a "radical zero-shot" approach in which the AI is given no codebook and is simply told to invent its own codes (in Spanish). We gave the AI context about the project.
- We found **251** causal links mentioned by the respondents
- Then we also auto-coded the sentiment of each link in order to show which contributions were "positive" (blue arrowheads) and which were "negative" (red arrowheads).

## Step 4: Answering research questions with Causal Map

- Once the coding was done, we used the filters in the app to create different maps that answered their research questions:

  - "What was the immediate impact on the respondents' lives because of gender discrimination?"
  - "What is the causal network from gender discrimination?"
  - "What are the most mentioned factors by the sources?"


desarrollo profesional (15)
educación formal (7)
mejoras profesionales (10)
responsabilidades laborales (13)
discriminación de género (6)
falta de experiencia laboral (7)
desempleo (4)
apoyo familiar (10)
motivación personal (9)
mejora en relaciones (7)

- We also used the ‘AI Answers’ feature to help us understand more about the interviews
  - This functionality allows you to ask questions about all the text in your file.
  - It is completely independent of causal coding. It will work just as well without causal coding.

## See what Javiera Cienfuegos, Senior Researcher of the evaluation project, has to say:

👥 "The type of questions that were asked "what causes what", were equally linked to methodological innovation. The results were able to portray how gender barriers are intertwined in domains ranging from higher STEM education to the performance of new professionals and technicians once they enter the labour market, reaching deeper explanations and social impact.”

## Qualia and data security

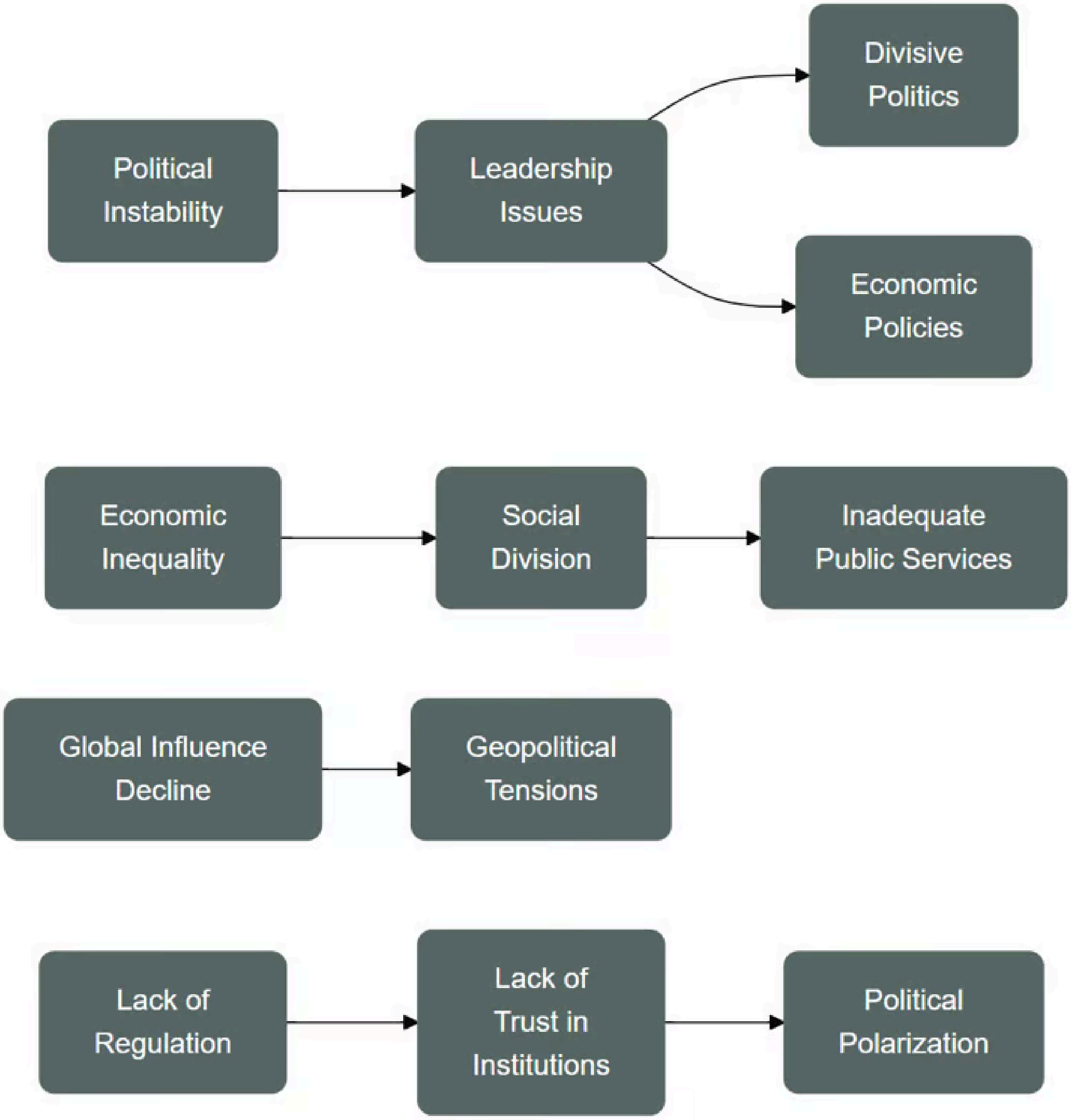
- Data security and compliance
- Our candid opinion is that although many clients are understandably extremely cautious about using AI, the risks are completely within the range of any other online data collection, e.g. questionnaires. A system is only as secure as its weakest link. For example if datasets are being shared by an online service like Google Drive, there is not much point having a Fort-Knox-level AI service ...
- For Qualia:
  - The transcription and Interviewer APIs are located in the USA, at openAI's servers. Data is not used for training. Data is retained there for 30 days for US compliance purposes.
  - Interview data is stored at a Heroku SQL database in the USA. Data is encrypted at rest and in transit. This is standard best practice. We have daily backups. While it is relatively easy to move the location of AI services it is quite difficult to move the location of database servers.
  - Clients are sometimes concerned about the AI data being temporarly stored in the US. However, so is just about everything else that happens on the internet .... If the client requires using AI services located say in the UK or EU, we could probably do that. But it is not obvious to me what would be gained. It is in theory possible to also provide AI services which are not retained for 1-30 days for compliance purposes. However this may require justification and could conceivably attract the attention of e.g. anti-terrorism agencies.
- **The Data Security Sceptic**
  - A client who is primarily concerned about data privacy, security compliance, and the risks associated with AI-powered interview tools.
  - Qualia's data security measures are on par with standard online data collection methods, with encryption both at rest and in transit in the SQL database.
  - Interview data is not used for AI training, addressing concerns about proprietary or sensitive information being used to improve AI models.
  - The temporary 30-day storage of data on US servers is comparable to most internet services and tools commonly used in research. There are good reasons for the data retention controls.
  - Daily backups ensure data integrity and protection against loss.

## Qualia asks about USA problems, again

# Feb 27, 2025 at EES

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.

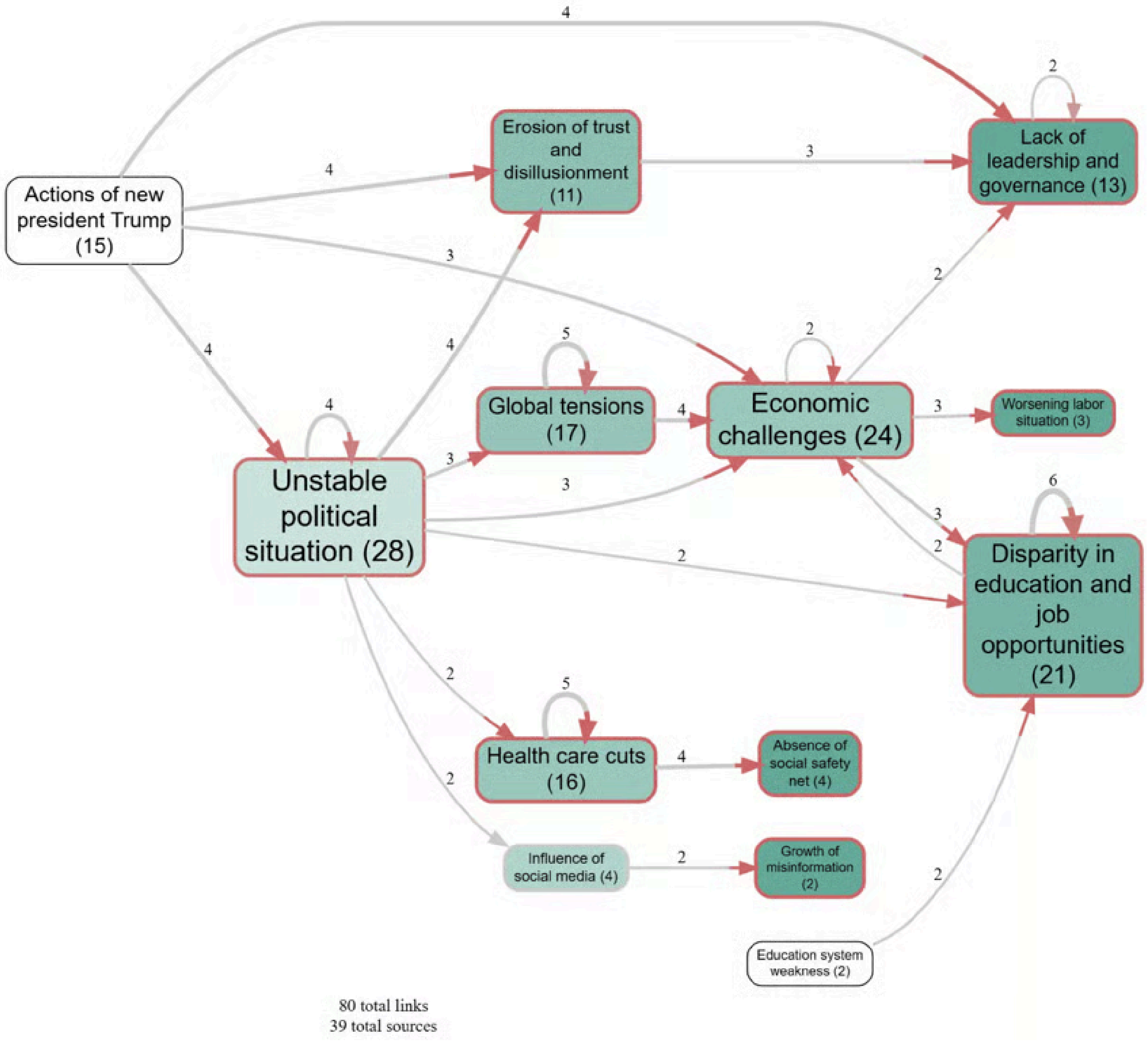
The Qualia platform provides an instant overview of the transcripts. For some reason, we didn't think to show it at the time, but I've pasted it in at the bottom of this post. Qualia also provided a simple causal map:



*Because this was a demo interview and many respondents only started it and only a few finished the conversation, we are not taking this analysis so seriously, it's just an example of the types of outputs you can get with the Overview Tab in QualiaInterviews — but although we can't make any claims to be doing fundamental social science here, the results are still worth a look.

The Overview in the Qualia Workspace app is just a simple hack which is basically like uploading all the transcripts to ChatGPT and saying "make sense of this please". We've already talked at length about the dangers of that: basically you are entrusting a whole load of evaluative judgements to a black-box AI, which is not only completely non-transparent but is cutting corners everywhere in the attempt to come to a plausible enough result as quickly and cheaply as possible.

A much better way is to break up the vague, high-level task into multiple simple, transparent ones, in this case, identifying all the causal claims in the transcripts, where someone said that one thing leads to or influences another, and aggregating them. The result looks like this:



Actions of new president Trump (15)
Erosion of trust and disillusionment (11)
Lack of leadership and governance (13)
Global tensions (17)
Economic challenges (24)
Worsening labor situation (3)
Unstable political situation (28)
Disparity in education and job opportunities (21)
Health care cuts (16)
Absence of social safety net (4)
Influence of social media (4)
Growth of misinformation (2)
Education system weakness (2)
4
4
3
2
3
2
4
4
5
4
4
2
3
3
3
3
2
2
6
2
5
4
2
2
2
80 total links
39 total sources

*A "Factor" is any box, including outcomes, drivers and things in between The map is filtered to show most important links and/or factors: many other links and factors are hidden Numbers on factors (boxes): number of mentions Sizes of factors (boxes): number of mentions Numbers on links: number of sources mentioning it Darker backgrounds: higher "Outcomeness": a bigger proportion of incoming links Deeper red arrowheads: the effect was more negative in significance/sentiment*

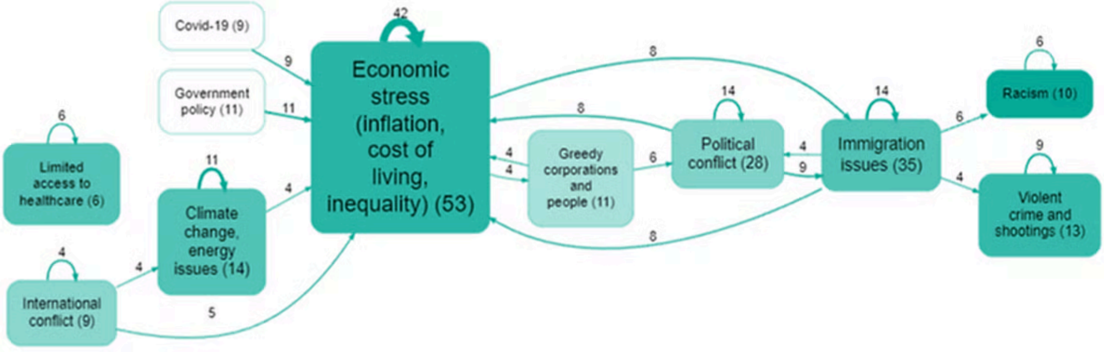
Some things to note:

- Many people mentioned Trump as a driver of changes (white background, positioned at left)
- Most frequently mentioned factor was "Unstable political situation", whose only significant driver was Trump's actions.
- We shouldn't fall into the "transitivity trap" of thinking that, because Trump is linked to Unstable political situation which is linked to Health care cuts that many or any individual sources told us about all the sections of this chain: the information for each section might have come from different sources (in fact, it mostly did).

We have done this type of interview several times before. Here is a map from 2023.

## EES 2023: Sharing our journey on AI's application in qualitative research

This was a completely different method and sample. The difference between these two maps has substantial face validity, but that is about all we can say at this point.


Covid-19 (9)
Government policy (11)
Limited access to healthcare (6)
Climate change, energy issues (14)
International conflict (9)
Economic stress (inflation, cost of living, inequality) (53)
Greedy corporations and people (11)
Political conflict (28)
Immigration issues (35)
Racism (10)
Violent crime and shootings (13)
42
9
11
4
6
11
4
4
5
8
8
4
4
6
14
4
9
14
8
6
6
4
9

---
Filename: usa-problems-merged. Citation coverage 53%: 343 of 653 total citations and 83 of 88 total sources are shown here.
Numbers on factors show source count.. Factor sizes show citation count. Darker factor colours show greater outcomeness.
Numbers on links show source count.
Zooming out to level 1 of the hierarchy. Auto clustering factors with granularity of 56% at level 1. Showing only factors with at least 6 sources. Showing only links with at least 4 sources.

# Technical annex

## The interview

The instructions for the AI interviewer are to conduct a fairly mechanical interview which simply asks the respondents to list the main problems facing the USA today, and for each problem, to ask for reasons and reasons for the reasons, and to ask how these are interconnected. It also asks the respondent to confirm the causal connections identified before finishing. It isn't a very chatty or explorative interview, but it works.

## Producing the maps

We asked our AI to code the transcripts one at a time and to use codes for each cause and effect which were already a little abstracted, using "concepts from a social science textbook" but with no more specific codebook. This resulted in 325 links using 458 factor labels with overlapping meaning, which we then clustered using a procedure detailed in a forthcoming publication.

## Instant overview of transcripts provided by Qualia (global, "black box" approach).

### Overview of Interview Transcripts

The interviews highlight a wide array of challenges currently facing the USA, with themes revolving around political, social, and economic issues. Common problems include divisive politics, leadership concerns, economic inequality, governance issues, and the influence of certain political figures like Donald Trump.

### Common Features:

- **Leadership and Political Instability**: Many respondents highlight issues with leadership and governance as pivotal problems. Concerns include the actions and influence of Donald Trump, new administrations, and an unstable government.
- **Economic Concerns**: Economic inequality, unemployment, and budget deficits are frequently mentioned, pointing towards systemic economic challenges.
- **Social Issues**: Polarization, racism, and loss of public services are recurrent themes, along with critiques on social justice and inequality.
- **Foreign Policy and Geopolitics**: Geopolitical tensions, particularly involving countries like Russia and China, and the USA's diminishing global influence are noted.
- **Environmental and Social Policies**: Climate change and health care are cited as critical problems needing more attention.

### Differences:

- While some respondents focused on specific figures, like Trump, and their impact on the national and international stage, others highlighted systemic issues such as the neoliberal economic model or the enduring influence of oligarchs.
- Perspectives vary on underlying causes, with some pointing to specific policy decisions and others referencing broader societal trends, such as populism or media influence.

## 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" (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 & 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.

This paper presents results from a proof-of-concept analogue study. We employed online workers as respondents, recruited via Amazon's MTurk platform (Shank, 2016). We decided to investigate respondents' ideas about problems facing the USA, 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 USA. As it does not encompass a specific intervention this theory is not an example of a program theory.

This interview guideline was implemented via an online interview "AI interviewer" called "Qualia", 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” 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 timepoints in September, October and November 2023, inviting approximately N=50 respondents each time. The data from the three timepoints was pooled.

## References

Ferretti (2023). *Hacking by the Prompt: Innovative Ways to Utilize ChatGPT for Evaluators*. https://doi.org/10.1002/ev.20557.

Jang, & Lukasiewicz (2023). *Consistency Analysis of ChatGPT*. https://doi.org/10.48550/arXiv.2303.06273.

## 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.

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 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:

1. **Inductive clustering**. First, we grouped the labels into clusters of similar labels using the hclust() function from the stats package of base R (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 weren't, the instructions were tweaked or rewritten and tested again to ensure quality and consistency.

# References

Team (2015). *R: A Language and Environment for Statistical Computing,*.

## Using AI interviewing - beware of bias

(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.

# References

Azzam (2023). *Artificial Intelligence and Validity*. https://doi.org/10.1002/ev.20565.

Head, Jasper, McConnachie, Raftree, & Higdon (2023). *Large Language Model Applications for Evaluation: Opportunities and Ethical Implications*. https://doi.org/10.1002/ev.20556.

Reid (2023). *Vision for an Equitable AI World: The Role of Evaluation and Evaluators to Incite Change*. https://doi.org/10.1002/ev.20559.

## 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.

See also: You have to tell the AI what game we are playing right now