CHAPTER

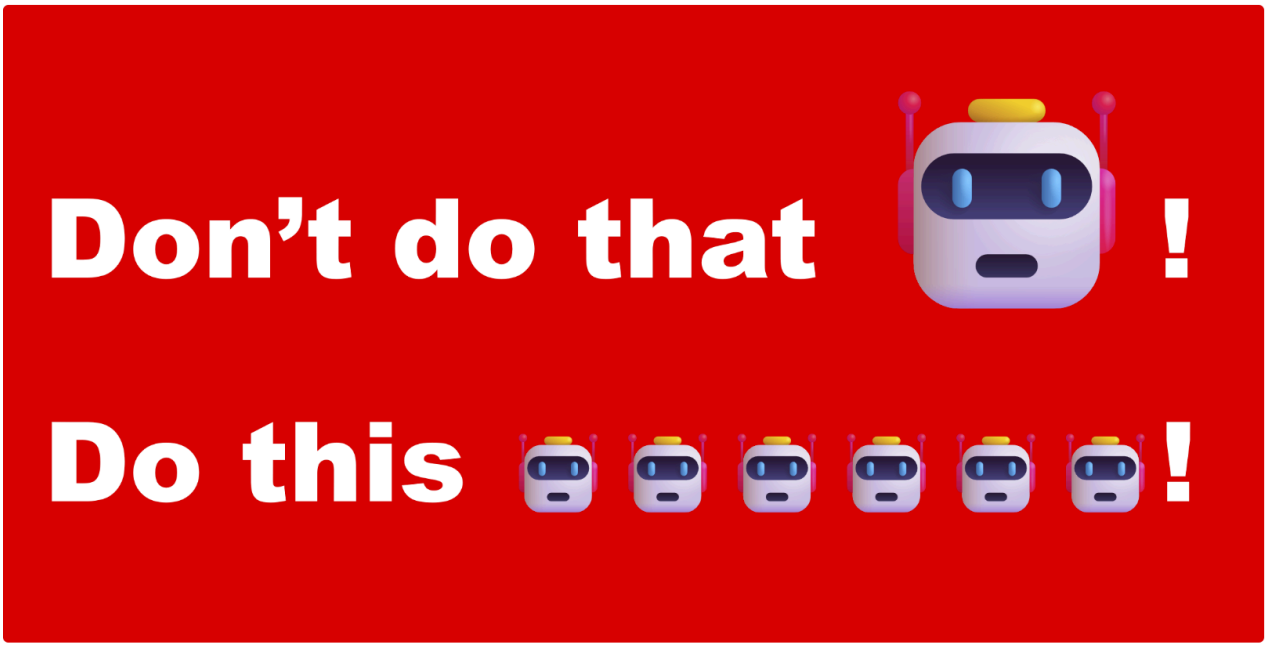
Improving rigour in the use of AI in social science

CONTENTS

Intro

How can we improve rigour and even reproducibility when using AI in social science? This chapter suggests some answers.

Just add rigour Three do’s and don’ts



Three do’s and don’ts when using AI for text analysis.

A lot of evaluation work is a kind of text analysis: processing reports, interview transcripts, etc. A bit like qualitative social science research. So this little piece is for evaluators in particular and (qualitative) social scientists in general.

How do we get from texts to evaluative judgements?

Recently many evaluators and researchers have been turning to AI to help.

BUT if you didn’t have a clear workflow from data to judgements *before* AI, don’t lean on the black box of the AI to cover that up. Here is my first set of Do’s and Don’ts. More soon.

1) DO Break up big, vague tasks into multiple smaller, clearer steps

Do	Don't
DO Break up complex, vague tasks into smaller steps which can be intersubjectively verified.	***DON'T*** Ask AI to make broad evaluative judgments (like "Is this good?")
DO Document your methodology so that you can explain step by step how you reached your conclusions in a way which anyone can check. No black boxes. Use the AI to speed up many simple tasks which you *could* have done yourself if you had the time.	***DON'T*** Trust the AI's explanations of how it reached its conclusions. AIs often create plausible-sounding but unreliable explanations after the fact. Normal AIs have very limited information about their inner processes
.	
DO Break up the data into pieces for AI analysis. Ideally run each piece as a separate prompt. Failing that, number each section and ask for a numbered, section-by-section answer, for example in a table.	***DON'T*** Give an AI large pieces of text and expect it will pay due attention to all of it. It will *claim* to have done, and may provide references to relevant passages, but attention is *expensive* and it is always trying to reduce that expense. If you let it, it will always try to skim read and jump to conclusions.
DO Use explicit, manual methods (Excel?!) to synthesise the results of the multiple separate tasks you gave the AI.	***DON'T*** Ask an AI to do maths for you, like adding up the number of positive or negative findings on a rubric. AIs are still terrible at maths.
Even worse, DON'T ask an AI to do *implicit* counting and comparison like "are there more positive or negative mentions of X in this report?"	

AIs excel at specific, well-defined tasks that can be verified intersubjectively, like rubrics. Most importantly they can answer lots of them, quickly.

"Intersubjectively verifiable" just means that most people will more or less agree on the answer most of the time.

- It creates transparency and allows others to verify your work.
- Clear instructions lead to more reliable results.
- If you can't check it, you can't trust it.

Example of an intersubjectively verifiable task:

> ✅ Does this paragraph mention water and sanitation?
>
> ✅ If so, are any recent changes mentioned?
>
> ✅ If so, do these sound like positive changes according to the interviewee?

Notice that here we've broken down a larger task into three smaller and simpler steps.

Examples of tasks which are *not* intersubjectively verifiable:

> ❌ Is the intervention described in this report efficient and effective?

Text needs breaking up into sections, judgements on efficiency and effectiveness need breaking down into pieces, e.g. using rubrics.

> ❌ What are the main themes in this document?

This is a very common question in qualitative research, but it's a terrible task to give to an AI without further details. What do we mean by a theme? Are we interested in economic aspects? Interpersonal aspects? How are the themes to be identified and refined? Here, a whole world of qualitative social science experience, skills and workflows (grounded theory, thematic analysis) have been bypassed in a single sentence.

> ❌ Summarise this document!

Yes, everyone does it. Evaluators do it. Schoolchildren do it. Pets will be doing it soon. As a quick time-saver for low-stakes tasks, it's very useful. But it's the vaguest, highest-level instruction, not a systematic analysis.

How do you break down a high-level judgement into a workflow of smaller tasks? Well isn't that what evaluation methods and qualitative research methods are for? Go read a book!

We're not saying you have to specify *in advance* exactly what methods you will use. That's a bit too positivistic. But you should at least document them as you go along and be prepared to defend them when your analysis is done. That's the untranslatable Nachvollziehbarkeit.

At Causal Map Ltd, we've found that highlighting and then aggregating causal links is a great and relatively generic path from text data to the brink of evaluative judgement.

In terms of how to implement your workflow technically, see this great contribution from Christopher Robert. At Causal Map, we're also working on ways to make workflows accessible. See how we currently use AI in Causal Map here.

This post is based on my recent contribution to the NLP-CoP Ethics & Governance Working Group, along with colleagues Niamh Barry, Elizabeth Long and Grace Lyn Higdon. In the next couple of weeks we'll look at two more do's and don'ts.

This post was originally published by Steve Powell on LinkedIn and has been republished here. See the original article here

Put down that thesaurus -- an open call to qualitative researchers

Trust the algorithm, not the AI

I often hear concerns about algorithms and AI, in everyday life as well as in evaluation, taking over our lives or making us submit to decisions made by machines.

The worry about losing control to machines is real, but we need to distinguish between different cases, and in particular between **using algorithms to make decisions** and **using AI to make decisions**, especially **evaluative decisions**. This is particularly relevant in the field of evaluation.

An algorithm is simply a set of explicit steps to make a decision or produce an output, usually expressed in code or clear language. Organizations have used such rule-based systems for decades.

Some different ways to make decisions

No algorithm: trust the human

The alternative (precursor) to algorithms is trusting humans to make decisions. This can be great if humans consider context and individual circumstances, what Scott calls "mētis," or local, practical, tacit knowledge, (Scott, 2020) but it can also lead to bias and corruption.

We can see **rubrics in evaluation** (King et al., 2013) as a kind of soft algorithm. We usually welcome rubrics because they make evaluation criteria more explicit, transparent, and less subject to the whims and unreliability of individuals.

Algorithms based on explicit criteria

Algorithms can help decide things like student admissions or loan approvals using clear steps (e.g., check age, if under 18 go to step 12, otherwise continue with step 5 ….). When implemented wisely, algorithms can improve fairness and consistency compared to human judgment alone.

Using statistical models

Some algorithms use statistical models to predict outcomes, like creditworthiness, by combining data such as age or location. A statistical model uses parameters like age or location each of which has shown to be associated with the outcome, which makes it somewhat transparent.

Both explicit and statistical algorithms can be criticized for bias, but at least they can be transparent if their rules are published. Problems arise when rules are hidden or people are discriminated against because of the groups they belong to.

In a more advanced statistical model we might find it increasingly hard to understand where the different parts of the formula come from: it might combine parameters in ways which for us seem meaningless and hard to justify but which are supposed to be associated with the outcome of interest. Opaque models can become what data scientist Cathy O'Neil calls 'Weapons of Math Destruction' (O'Neil, 2017).

Machine learning

Machine learning is a subset of artificial intelligence where systems learn from data to identify patterns and make decisions or predictions, from "is this a picture of a cat" to "should we approve this person's application" often without being explicitly programmed with step-by-step rules. Instead of following a predefined algorithm, ML models develop their own 'rules' (which are often opaque to humans) based on the data they are trained on. Unlike generative AI, you can't chat with a machine learning model, you give it input in a fixed format (say, a picture) and get a fixed output, e.g. yes/no.

Sandra Seitamaa https://unsplash.com/photos/a-dog-and-a-cat-sitting-on-a-couch-Y45fzr5p3ug

In the extreme case we might have an algorithm based on machine learning (a form of AI, but not generative AI), where perhaps a neural network has been trained to distinguish desirable from undesirable candidates in just the same way you can train it to recognise a cat or distinguish a cat from a dog. Machine learning can be used to make decisions without clear formulas or rules. The process becomes a “black box,” where we input data and trust the output without understanding how the decision was made.

Generative AI

Generative AI is a type of artificial intelligence that can create new and original content, such as text, images, audio, or code, after having learning patterns and structures from large datasets.

These models don't just classify or predict, but generate novel outputs based on the input they receive, for example, continuing a conversation or answering a question.

The most extreme case is using generative AI for evaluative decisions without clear criteria (using it as a big black box): simply asking the AI, for example:

- is this program component effective?
- should this client get a loan?

Conclusion: make good use of algorithms

People often misunderstand algorithms, which can provide explicit and transparent decision-making. The real concern is not so much the use of algorithms but the shift toward the use of machine learning and generative AI, where the decision-making process becomes less and less transparent.

> Using AI in decision-making can be worrying not because it uses algorithms but because it *doesn't*.

What's your positionality, robot

> Imagine two researchers coding interviews about the cost of living. One grew up in a wealthy family, while the other experienced poverty first-hand. Their backgrounds will certainly influence how they code.

Nowadays, people are using AI for text analysis. Many of us worry about AI's "**hidden biases**". What to do about that?

Often there is no such thing as being objective, but at least we humans can be explicit about our positionality, our background and motivations, how this might affect our work, and how this relates to the positionality of our audience.

What about with an AI?

You can ask an AI to explain or reflect on its positionality and it will certainly give a plausible response, but remember that an AI has in fact very little insight into its own workings. Perhaps it will suggest always being aware that it was trained on a specific set of data which is not representative of the whole of humankind.

In any case the criticism that AI training data is not "representative" misses the point. Even if the training data had somehow been representative of the whole of humankind, that wouldn't make it "objective". It would simply reflect humanity right now, with all our quirks, biases and blind-spots. It wouldn't mean we don't have to worry about AI positionality or bias any more. It wouldn't (of course) mean we could rest assured that everything it does will be morally impeccable.

What's most unsettling about working with AI is not that secretly it's a bad person. The problem is that secretly it isn't any person at all. Even if it (sometimes) sounds like one.

A suggestion

A better suggestion is to be **more explicit about positionality in writing prompts and constructing AI research workflows**. Here is a very humble idea about how to start this experiment.

A simple example: I can tell my AI:

> When working, implicitly adopt the position of a middle-class white British left-leaning male researcher writing for a typical reader of LinkedIn. Don't make a big deal of this, but it might be helpful to know what your background is supposed to be before you start work.

And we can start to add variants of the kind of procedures which we humans might use when trying to address positionality:

In my AI workflow, I can then give another AI the same task but with a different starting position, and then perhaps ask a third AI (or a human!) to compare and contrast the differences. That also crosses over into ensemble approaches.

Of course, adding a phrase like “middle-class white British left-leaning male researcher” does not mean the AI will suddenly have all the relevant memories and experiences or really behave exactly like such a person. It’s just a fragment of what we mean by "positionality”. But *it’s a start*.

Have you been experimenting with this kind of approach? We’d like to hear from you!

Footnotes

At Causal Map Ltd, we’re working on an app called Workflows to make AI work more transparent and reproducible.

We’ve found that highlighting and then aggregating causal links is a great and relatively generic path to make sense of text at scale.

In terms of how to implement your workflow technically, see this great contribution from Christopher Robert.

See how we currently use AI in Causal Map here.

This post is based on my recent contribution to the NLP-CoP Ethics & Governance Working Group, along with colleagues Niamh Barry, Elizabeth Long and Grace Lyn Higdon.

This post was originally published by Steve Powell on LinkedIn and has been republished here. See the original article here

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

It's strange how often this happens:

Humans are discussing some task, and one of them turns to an LLM to see how it would carry that task out. Sometimes the results are disappointing or seem to demonstrate that LLMs are, after all, stupid or limited.

Normand Peladeau, on the QUAL-SOFTWARE mailing list 7/11/2025, reports having tried just that with the famous (or infamous) Sokal Hoax text. He asked different LLMs whether he should accept a paper proposal for a philosophy of science conference. The proposal was the first two paragraphs of the Sokal Hoax text. (Spoiler: the leading models like GPT-5 recognised the text anyway; some of the others seemed to fall for it.)

But: Is that enough background? Is a simple sentence enough to bring the LLM up to speed with the crucial background information *what game are we playing here*?

Don't forget that the LLM does (mostly) not know who you are or what you are expecting or what kind of conversation you were just having. Perhaps you are expecting something humorous, or informative? Perhaps you want ideas to start the next chapter of your novel? Perhaps you just want the LLM to respond as many (over-)educated humans might do: and after all, **actual humans did fall for the hoax**!

To be a meaningful and useful test which might extend our understanding of the strengths and weaknesses of LLMs, we should make sure we explicitly add the extra context of **what kind of game are we playing here**. Is it a serious review? What do we consider the role of a serious reviewer? What are we looking for?

So maybe our conclusion should be: you can't expect LLMs to guess what you are thinking, out-of-the-box. I don't actually know how well different LLMs would perform if we gave a more precise contextual description before setting the task; after all, we all love that warm feeling of Schadenfreude when an LLM fails at something, but the feeling is even warmer if the test was a fair one!

We have this kind of problem often when helping clients write interview instructions for our AI interviewing platform, QualiaInterviews.

Clients know they could themselves lead the interview well because they have all kinds of background information and expectations, much of it only half-conscious, from the general style of interview they expect, how much this particular interviewee can be pushed, how much warm-up chat they might need or expect, what are the most important research aims, which themes can be skipped, and so on. Clients might get frustrated when the AI fails to have read their minds when leading an interview, but they have to ask themselves: what additional information would even a gifted and experienced human interviewer need if they knew nothing at all about the context, the client or any of the background? I think something similar applies in the case of Normand's very interesting experiment.

