This is a transcript of a seminar from our series **Innovative Approaches to Research in Equity, Diversity & Inclusion** hosted by the EDI Caucus and held online on 13 May 2025, chaired by Fenella Watson, EDICa’s Senior Project Support Officer. The title was Using AI to Transform EDI Research.

The presenters were Dr Lukas Kikuchi and Sean Greaves of the Autonomy Institute.

Headings have been used to facilitate navigation in this document, and the text has been lightly edited for readability. The recording and the slides can be viewed here:

# Introduction

Good afternoon. Welcome to the second seminar in the EDI Caucus' Research Methods series. My name is Fenella Watson and I'm the senior project support officer on the Caucus and I'll introduce Lukas and Sean our speakers after this introduction. We have about 20 minutes of presentation plus a demo of the Scribe tool that they're going to show us and followed by a Q&A and discussion.

## About the EDI Caucus

For those who are new to the EDI Caucus - that's EDI meaning equality diversity and inclusion - it is a project and initiative that was funded by UKRI with additional support from the British Academy. It started in early 2023 and is due to finish in 2025 however we're looking for an extension and of course we're looking for further funding to continue our research. Some of the key actions of EDICa are to research the evidence base for inclusive careers across the UK's research and innovation systems, to remove barriers experienced by marginalised groups and support them to access and thrive in chosen careers, to conduct interdisciplinary research, to create national and international communities of practice, and to use a co-design approach with stakeholders and people with lived experience.

This is the second in the Research Methods seminar series that we've started. In an increasingly diverse and interconnected world the need for rigorous research methodologies to examine equality (or equity), diversity, and inclusion has never been greater. The EDI Research Methods seminar series is dedicated to equipping researchers and practitioners with the tools, frameworks and approaches necessary to critically analyse and address EDI issues across various domains. Here you can see some upcoming seminars. We have about one per month. We've got 11th of June - Social Network Analysis; 2nd of July – Machine Learning and Gender; we take a break for August just for the holidays; 4th of September we have Literature Review and Black Women Academics' Experience; 2nd of October – Co-production in the Context of Neurodiversity. We take a break for November, which is when EDICa is having our project symposium or conference, which will be hybrid so you can join us online for that. Whether we have one in December is ... we're still working on that one.

We do record these seminars and we do share them on our YouTube channel and on our website for people to reference. We also, where we are allowed, put the slides there for people to reference for accessibility. Our website is supposed to be a resource tool - it's a library of resources on equity and research and innovation careers. We continue to work on that and to add to it. This is not just for EDICa's own research but also sharing other good practice and research conducted by other researchers.

And now to today's main event: Using AI (that's artificial intelligence) to Transform EDI Research. We have two speakers from the Autonomy Institute which is an independent progressive research organisation based in London. They've worked on the concepts such as the 4-day working week and the universal basic income. Autonomy are co-investigators of the EDI Caucus and they particularly work in workstream 3 which is looking at enabling workspaces. So without further ado, I will pass over to Lukas Kikuchi and Sean Greaves from the Autonomy Institute who will introduce themselves.

# Presentation

[Lukas:] I'm Lukas Kikuchi. I am the head of the Autonomy Data Unit at the Autonomy Institute. The Autonomy Institute is a progressive research institute. We produce policy, we also support the general progressive sector for other organisations in particular. My team - the data team - we do data support work – bespoke technical work for other progressive organisations. We also produce tools for both our in-house policy team and consultancies and also for external organisations. One instance of this is what we're going to talk about in our presentation. Sean, do you want introduce yourself?

[Sean:] Hi everyone, great to be with you today. I'm Sean Greaves. I'm a researcher at Autonomy. I work in the data team with Lukas. More specifically, I lead much of the prototyping for a number of Autonomy software tools, and specifically the ones that are using LLMs and some of the latest AI technologies. Prior to Autonomy, I was working in machine learning solutions engineering in the tech sector. Companies including IBM.

[Lukas:] Today we're going to be talking about Scribe, which is an in-house tool that we've developed. It's not a tool that you can find online right now. It doesn't have a website, although we will present a sketch of a demo at the end, which will show some of the capabilities of Scribe. It's essentially a methodology or a tool that we've applied a few times inside Autonomy, and it's an AI tool to deal with large sets of free-form text survey responses. Before I go into Scribe, I'm just going to briefly go over a bit about our organisation.

## About the Autonomy Institute

Autonomy is a not for-profit, where we do independent research. We do both our own research and we work with other progressive organisations and some political parties. We are funded by a mix of core funding foundations, and also doing progressive consulting work and other kinds of client work. We're progressive; our main focus is broadly freedom, equality and sustainability. We do work on the shorter working week; we do work on the universal basic income; we also have done work in climate change. Recently we've also engaged with some of the discourse on the emergent far right. Finally we like to say that we walk the walk. We have forward-thinking workplace policies. We obviously produce policy on the shorter working week but we also have a shorter working week – 4 day week.

Our work can be very broadly split up into three. We do research; a lot of the research we do has a data quantitative component – that's the purview of my team. If you do a policy report on universal basic income – the economic analysis, the micro simulations, will be done by my team to support the policy recommendations we make. We also do strategy work. We consult for different branches of government. We've done work on labour skill shortages, on green transition; we work with investigative journalists and trade unions to both enhance their practice and also shape their strategic aims. We also produce intelligence. This is linked to the strategy. We produce intelligence by doing a lot of web scraping. To some extent, the tool we're about to present here is a manifestation of that as well. We want to be able to source opinions, source qualitative information from a wide range of sources, and we've often deployed AI and other tech stacks to get a lot of intel on various political issues.

## Autonomy’s use of AI

And as I mentioned, we use AI a lot in our practice. We've used it even before - maybe two years ago, when Chat came out, but the game has changed quite a lot since then because of ChatGPT. It's been semi-revolutionised. That has also meant that the principles by which we do approach AI has had to adapt as well. Primarily, because of the power of AI, but also the potential pitfalls of AI. And when I say AI, in this sense, I do refer to the recent boom of large language models, generative AI… the thing you're talking to on ChatGPT basically. We've developed an approach to AI that tries to exploit all its capabilities whilst also remaining cautious of the AI hype bubble.

The three principles that we've developed is, firstly, the idea of lossless processing. If you give a long text to Chat and ask it to summarise it you'll get a summary, and often it's quite good, but if you want to present that to other people, if you want to present that to clients, you want to do it in such a way that you have ease of access to the original source material. That's what we mean by lossless – we can transform some of the data using AI, but we want to do it in such a way that we don't lose the original source data. And as much as possible, we don't transform at all, we just use AI to organise and categorise the original data.

Secondly we want to use AI to enhance the human researcher. Mainly, AI can be used as a cost cutting tool, but we're a bit more interested in using AI to do things that weren't possible before. And what we're about to talk about today, which is using AI to look at large scale survey processing, you could say that's a kind of cost cutting tool, but really what we're talking about is being able to do a thing that, primarily due to cost, has been unfeasible. Looking at very, very large scale free text surveys. We're all about enhancing the human researcher in a specific research stream using AI, but we don't want to use AI to do the research for us.

Finally, we want to leverage mediocre computing. This phrase “mediocre computing” refers to the fact that, whilst ChatGPT or other LLMs can be very clever, perhaps the best use of AI is the fact that it can reliably be quite mediocre. We don't necessarily want to use it for very complex tasks or tasks that require complex reasoning or judgment. Rather, we are more interested in exploring how we can use AI to do quite simple tasks, but at scale – slight rephrases, rephrasings of sentences, summarising. Those are the kind of things that are not, on some level, very complex. But to be able to summarise millions of texts in a minute is quite an impressive thing – that large deployment of what we call mediocre computing.

## What is Scribe?

With that general introduction out of the way, I want to go on to the main topic, which is large scale processing of survey data using Scribe. What is Scribe? Scribe is not an API or service that we've exposed publicly. It's rather a methodology and an in-house tool that we've developed at Autonomy for our own qualitative analysis of survey data. So far, we've done a number of things with it; we've organised large clusters, we organised and clustered large volumes of unstructured text. These are surveys, interviews and consultations. And within them we then use Scribe to track emergent themes from the various voices inside those corpuses of text and then we try to quantify the qualitative patterns, whilst also preserving access to the original source text.

With Scribe we don't try to replace the human researcher, we don't just pop in a big Excel file of survey responses and see what comes out the other end, we rather use that as like a first draft of analysis, or just a first reorganisation of the survey responses, and then a human researcher (that being one of us) will look at the survey responses, maybe adjust some things that were misclassified. In a way, we use Scribe to get 90% of the way in a lot of the analysis we do and then the last 10% are the most important of doing the finishing touches. And recently we've been trying to develop Scribe into a client-facing service to support other organisations.

This presentation and maybe the discussion following will be good feedback to that end. Mainly we've been using it in combination with our shorter working week work. We do a lot of consultancy for the shorter working week and part of that process is to do exit surveys. We also do a lot of survey work in general.

Originally all of our survey work was completely categorical – by which I mean “how did you enjoy the 4 day working a week?”. I really liked it, I liked it, neutral, etc. But using Scribe we've been able to look at more higher dimensional survey data which is free-form text.

I'm now going to get slightly technical. I'm going to go through the actual process of Scribe and it's not overly complicated, although it does use these kind of frontier technologies. The main two being embeddings and LLMs, which I will try to describe a little bit. I've been asked to make it clear what tools we're using, but for that I'll just give an overview of what Scribe entails.

There's essentially just three steps to Scribe. We take a potentially quite long text response and we split it up into several smaller responses. I'll give an example of what that means; if you give the opportunity to a survey respondent to write several paragraphs of text in answer to a question, you might get, in effect, several responses. You might want to split these into several smaller ones. This step one is optional, and sometimes you want to do both where you don't split it up and you split it up and see what the difference is.

In step two, we then cluster or group/categorise similar responses by semantic proximity. This means that if two responses are similar in what they're actually saying, we cluster them into one group. The key tech we use to do this is by converting the text to what's called an “embedding”, which is a sort of mathematical representation of the meaning of the text.

Finally, once we've grouped or clustered all the responses, we then summarise each cluster using an LLM. This is mainly to give a bird's eye view of what the content of the survey is.

The first part of the Scribe methodology is to split up a survey response. For example, here on the left hand side, I have an example of a survey question. "What has the shorter working week changed in your working life?" and then you might get a response that has a lot of different things. "I'm much more focused when at work... it helps me with catching up. I love the shorter working week, I feel that I get a lot more work done in time and business now versus before the trial" etc. Basically it's a semantically quite complex response. It has a lot of different things. One thing we do sometimes is we then split this up into several smaller ones. We split it up in such a way that each bullet point here is saying one thing. "I am much more focused at work during my working hours." "I have to be more diligent with my time management." "I strongly appreciate the shorter working week." "I find that I'm more productive now". This splitting will happen in such a way that the original response is still being kept but we identified what the individual responses are and this happens via the LLM.

You can try this if you have a ChatGPT subscription or account. You can paste this lefthand text here into ChatGPT. You have to work on the prompt a little bit, but if you write something along the lines of "can you look at this text response and try to split it up in bullet points of the individual things that this respondent is saying", you'll get something like this.

Once we have all of these responses, we use something called embeddings to categorise and cluster these responses. In this diagram, all these dots here on the right hand side are meant to be survey responses. Here is one from a shorter working week exit survey. Each dot is one split response. For example one of these dots could be "I'm much more focused in my work hours due to decreased working week". And the abstract space here that you see is not an actual distance as such. Each of these dots is a response and the distance between any of these two points is an indication of the semantic similarity of the responses. If two dots aka responses are close to each other, that means they roughly contain the same semantic content. The tool we use to do this is called "embeddings". It's a way to mathematically represent texts such that you can compare them in their semantic similarity.

What this allows us to do is clustering. Here, the colours of each dot is an automatic clustering algorithm that we've applied to group these dots into clusters of similar types of responses. And this is not always going to be foolproof. It might have to be adjusted and changed but it gets you 90% of the way into actually identifying what are people saying. It gets you from about a thousand or 10,000 individual text responses into maybe 12 or 15 or so groups that you can then look at. Groups which say roughly the same things.

To explain a little bit more what I mean by semantic similarity: two sentences that might be called semantically identical would be "she gave the book to her friend" and "her friend received the book from her". This contains almost identical semantic content. Two sentences can be semantically similar if it's slightly different, but it's roughly saying the same thing. "She handed over a book as a gift to her friend" is roughly the same as "her friend ended up with the book she gave". Although of course there's difference there, it entails different things, but you might group these things as the same thing basically. And this is essentially the kind of procedure we've automated to some extent.

Here I've zoomed into one of the clusters in the survey responses. These are responses to a question thinking about your local community. "What is the biggest change you'd like to see in the next year or two?" and here are some of the split responses that we found. We've had respondents asking for improved road safety measures, more spent on road improvement, improved infrastructure, improved infrastructure, better roads, etc. And as you can see, these are not exactly the same, but they're broadly about infrastructure. Potholes, road safety, but also broadband, better roads, etc. there's a thematic kind of continuity in this cluster and we've seen that in this automated process of using embeddings and clustering, we identified these, but there might be examples here that we might want to pick out – for example, broadband. I don't know if we want to... I mean it is an infrastructure concern, but maybe we want to have a specific cluster just for road safety, and things like that.

It's something that you might want to adjust later on, but it gets you 90% of the way in terms of actually doing the groupings of the free-form text. And then you might take all of the free-form text responses and then you summarise them. Using the LLM we then drive a summary. Once we've produced a spreadsheet of all these responses, the human researcher can then read the summaries, look through the individual responses, and then adjust accordingly. At the end of this, we get loads of the different clusters with a summary of each cluster. As Sean will show, there's other statistics as well – there's demographic data, potentially. You can make bar charts of how many people are in this cluster, how many people are in this cluster, etc.

## Use cases

Finally, I outlined some potential use cases for this kind of tech. Free-form text is of course richer than categorical surveys and primarily they allow for unknown unknowns. Categorical surveys might lead the question. They'll give you prescribed, potential possible answers but, if you open it up and allow people to write in the text box, you might get things that you didn't take into account. But if you do allow for free-form text, you might get reams and reams of text and you might not have the capacity to read through them. Scribe makes this scalable, because if you do ask for free-form text in the thousands or tens of thousands then you will get redundancies in terms of the fact that people might say the same things but phrase it differently. Scribe allows you to sift through this and see what are the unique insights and what we call, in this case, the intersectional insights. Potentially Scribe may also, with some adjustments of the methodology, perhaps, help us to surface marginalised voices.

Because we have the LLM that can read each individual response, it allows us to do a search for the responses themselves. There might be very underrepresented or minority perspectives, but if something sticks out then we can use LLM to sift through them. This can be combined with demographic data.

This is a cost or productivity tool to some extent because it can lead to faster feedback loops for engagement. We can reduce analysis times from weeks to days or hours; we can have a continuing process of survey sourcing and processing. We haven't applied it for EDI as such, but I think with what we're presenting, I think it can be adapted. I will hand over to Sean for him to go through the demo, which may make some of what I said a bit more concrete.

# Demonstration of Scribe

[Sean:] Thanks Lukas. I'm going to step through a case study that we have applied Scribe to, to give you a sense of some of the ideas, design decisions, and capabilities that Lukas was just talking through. In this case study we're going to use Scribe to investigate some data that came out of surveys from a shorter working week trial. As we mentioned, at Autonomy we run lots of shorter working week trials for a range of companies that are curious about testing this new policy. We work with a range of different companies on this. Some quite small in size – maybe 10 employees – up to much larger organisations that could have hundreds or potentially thousands of participants engaging in the trial who we will be getting feedback from at the end of the trial.

A bit more on the context in this sample – it's the results of a survey of four questions that we've put to a group of employees. We've had between 200 and 250 responses to each question – around 1000 responses. The question prompts an unstructured response; it's not a multiple choice; it is encouraging the respondents to provide free-form text. Obviously for the people who are reviewing these unstructured text answers, that's potentially quite a lot of work, if it is a thousand.

When we come to the end of a trial and we put a survey to a group of employees to try and assess how that trial has gone, we package up the data from that survey into a report. With a thousand responses that's going to take quite a long time to read, and that gets more and more challenging as the number of responses increases. It's also the case that once our researchers have read through all the responses, that's not going to be the only time that they look at those responses. They can't just go and write the report straight away. It's likely they will be doing multiple passes over the data as they attempt to taxonomise it to find the key findings, the key things that have been discussed. Again that process becomes more lengthy, more challenging, the more responses you have.

That's where Scribe comes in. Scribe is essentially where we take all of these responses and we cluster them. It makes it much more easy to find common themes, common points of discussion that can help our researchers when they are investigating all of this data and writing it up. In the interface, if I click up here in the config bar, we can see these are the four questions that made up the survey. The first question that we asked the respondents was "what has the shorter working week changed in your work life?" The second question is "how has the shorter working week impacted your personal life?" And what we can see is that when I select these questions, in the semantic space we've got all these dots appearing; every single dot represents a unique statement that was made by a respondent. We can see that for this question, we had 262 employees providing unique responses to the question.

As Lukas was mentioning, before we visualise this data in Scribe, what we're actually doing with each of the responses someone provides is splitting them up into unique statements. Because this makes it much easier to parse and cluster and find common points of discussion – particularly for cases where someone has provided a very long answer with lots of detail in it. They may have multiple points they wish to raise around how the shorter working week has impacted their personal life. if we want to read these responses, we can navigate around this window, and if we highlight over the responses, we can see in the bottom window it shows the content of that response. Every dot represents the response from one person. We can see for this one that I'm highlighting over it says in the bottom corner for the response the original data. I'll just read the first couple of lines, the respondent has said "the shorter working week has massively impacted their lives, they've been able to take hospital and doctor appointments on their rest day, there's no impact on work, allows them to spend quality time with their children, and doing the school drop and pickup, which they've never been able to do before.”

We can see already that in just the first couple of lines that there's multiple points made. What we've used Scribe to do is to parse out these individual statements that have been made. One of the statements that Scribe has parsed out reads "the shorter working week has had a great impact on my life as it allows me to make doctors and hospital appointments on my rest day without interfering with work". We can see there that this individual statement allows that point to be understood and to be arranged around other similar points that highlight the fact that the shorter working week enhances appointment scheduling.

We can also see below the statement that there is metadata as well. We get a sense of the role of the person who made the statement, gender, ethnicity, disability. Again this can be really useful for understanding across different groups what is more of a priority and helping our researchers to be more targeted in understanding how the trial has gone for different groups within the workforce.

If we explore around this window, we can explore some of the other clusters. This cluster over here highlighted in green, if I read the statement it says "I'm now able to dedicate more time at the gym". If we look over in the topics bar, we can see that every point is part of a cluster that is labelled as the survey reveals increased exercise opportunities. This is everyone that has written something about how taking part in the shorter working week has allowed them to exercise more. What we can do is get some stats on the number of respondents that are in this cluster. We can see that there are 25 respondents who provided an answer mentioning the increased exercise opportunities – about 9.5% of respondents. If we go down we can see the individual statements. If I zoom in on this cluster and hover over these points, you can see some of the different dots lighting up.

Again what we think is quite important with Scribe, and particularly when we're using AI models, is to give the user a sense of how data is being rearranged by the models. By being able to see which dots correspond to each response or statement you get a sense of what semantic space means and build a bit of intuition. As Lukas mentioned earlier, it is quite an abstract thing we're doing, taking all of these responses and arranging them into 2D and into clusters. It sounds quite abstract at first, so it is quite useful being able to have multiple ways of understanding how similar concepts or similar ways of saying things are arranged.

At the bottom here, one other thing we can do with Scribe is to link it up with a local database. This is useful when our researchers are potentially considering, for some of the findings they're seeing in the data, how this relates to reports that have been written before. In the further reading section, we've taken the whole of our organisation's publication catalogue and searched it for passages from reports that are relevant or correlate with the individual topics found in the data. In this case, because this cluster is talking about participants who've experienced "increased opportunities to exercise as a result the shorter working week", it's found this report we previously did on the shorter working week in Iceland. If I look here, an excerpt from that report, we can see someone is discussing having more time to go out and do exercise. This can be useful for the different niche topics, if we are looking to quickly see and find relevant publications or research we've done before that is relevant or agrees with that finding.

# Question & Answer

[Fenella:] We have a question: "can you provide any examples from your work of how this approach has helped to amplify responses from marginalised groups?"
I've had the advantage of being able to look at this demo in advance. So my other question was whether we could share a link to the demo so people can explore it further. When I was playing around with it there was the filtering for responses from disabled people, for example.

[Lukas:] Sean, go ahead and paste it in the chat. [<https://static.autonomy.work/adu/public/scribe-app/>]

We have two main ways. One of which is just an idea; it's something we've been toying with, so we haven't done it in practice. When we do survey responses like this, we have demographic or metadata in our responses. We can look at gender, ethnicity, disabilities, etc. We did a follow-up survey on the 4 day week a year on. [<https://autonomy.work/wp-content/uploads/2024/02/making-it-stick_-1.pdf>] We wrote a report based on this, and we used the Scribe methodology to look at all the survey responses, and then we also zoomed into different minority demographics. All the clusterings and things like that can be used in context with demographic data to look at what minority voices are saying. We looked at things like disability and gender composition, etc.

[Sean:] Let me demo this. Relating to EDI, one thing we can do is filter all the responses based on metadata provided about participants. For instance, I could select gender, I could select ethnicity, and we could investigate. These are all the responses for women of colour within the [responses]. We can look at certain points here. For instance, this is a small cluster of blue points. If I go in and click, we can see that 45% of the group of employees provided responses within this topic. It does allow us, at a very large scale, to filter, to zoom in and understand how different groups within the workforce are responding. Again, for our researchers that can be helpful because there are hundreds of responses and at first it can be very difficult to filter through and see a fine-grained analysis.

[Fenella:] That sounds like it answers another question which is "can the client choose what demographics they want to measure and expand beyond gender, ethnicity, and disabilities?" Is that depending on what metadata questions were in the survey initially?

[Lukas:] It is, yes. Basically, you should think of this as like the thing that Scribe processes. A general survey will have a gender selection box, other metadata and things like that, but then these can be combined. You might have a survey that has free-form text parts and then categorical parts, and the categorical parts then become metadata that then we can use to do other kinds of statistics.

[Fenella:] Do you have the option to expand on the category? Instead of saying "person of colour" it's the whole range of ethnicities that the census might use, for example. This is just depending on the design of the initial survey?

[Lukas:] Yeah. There's two things. Of course, we're not limited to specific categories, that's just completely up to how you design the survey. If you have more options for ethnicity then that's fine. But you can also think of Scribe as a way to have fluid categories in itself. In the ethnicity box, for example, you might not have categorical choices; you might have a little one-line text box where you write [your self-described ethnicity]. And then Scribe can be used to identify the unique kinds of responses in that specific category. You then get something that looks like metadata. That's another application.

[Fenella:] "What LLM are you using and what has it been trained upon, and given the known biases of, for example, ChatGPT and friends, are you not concerned about those biases creeping in?"

[Lukas:] The answer is yes. The LLMs we are using- we're a bit agnostic with respect to them, in principle that is. The methodology itself does not rely on a specific LLM, but of course the results will vary depending on it. We've been using a variety of ones. We've obviously been using the OpenAI ones – so that's GPT-4o, GPT-4o mini there's also Claude 3.5. We've also used a bit of Gemini. There's no way to avoid bias on some level – either because there's a human researcher dealing with these, so that's biased, and then of course there's ChatGPT, which obviously is biased. But one way we try to do this is we don't ask GPT to make subjective judgments. This is part of the principle that I discussed briefly in the beginning, which is that of mediocre intelligence. We treat ChatGPT as something that has mediocre intelligence or only rely on its ability to do very mediocre tasks- summaries and things like that. And that does not avoid bias, or at least it doesn't avoid the risk of bias, but that's also why, in the way we set up the methodology, the way we use it, it's not so much that we're throwing a big survey at the AI and asking it to do all the work, it's rather you should see Scribe as if you get a thousand responses and you get them in a big Excel file, it's going to take a very long time to read through them all, but you should still read through them all. When you get them in the Excel format, it's going to be in a random order – probably in a chronological order. Scribe lets you sort and organise these into more broadly thematic categories, but then it is still – in proper usage of this tool – you should still look through the responses and then see if the categories match to what you actually want to achieve. In some sense we're just offloading the bias onto the human researcher, but that's kind of unavoidable.

[Fenella:] Let me move on to another question while we still have some time. "Will it be a lot to ask the accuracy of the model?"

[Lukas:] In some sense, no. You'll never get full accuracy. For example, there will be misclassifications in the clustering, but the reason why it works is because it's largely accurate, and we're not asking too much of the AI. We're presenting the data to the researchers in such a way that it's encouraged that you make the adjustments necessary for any errors.

[Fenella:] "I like the use of linking publications" this person says, "to the identified topics. How do you choose the publications? Are they only the resources you have selected or written? And is there scope for expanding the publications it links to?"

[Lukas:] Yeah, in this case it's just our reports. That was mostly because we were interested in seeing what are the links to our corpus of work, but you can throw anything in there. Any body of work that you do, you can do that kind of automatic linking.

[Fenella:] The last one that we haven't got to is "Why your tool and why not ChatGPT and how is your tool different? Is it just that we can have the original data?"

[Lukas:] That's a good question. There's variety of reasons. The first is you want to use AI in a controlled setting. If you use ChatGPT you're going to have to use it once for every question or you just throw all the survey at the same time. We prefer having a pipeline where we send individual questions to the AI with the right context. It just gives us more control to calibrate the procedure. The main thing is that there's a context limit to ChatGPT – which means there's a limit to how much text you can give it. The approach of summarising things won't scale infinitely. The third thing is that part of what we're doing is the automatic clustering based on semantic similarity, and that is not something that ChatGPT does. Of course, if you give ChatGPT like 100 responses and ask "can you group these based on semantic similarity" it will do it, but that's only because ChatGPT will do anything you ask of it. We've built the pipeline to have a bit more control and also quality assurance of the actual thing we're doing. It’s also so that we can see original data. It gives us control of how we present the data and making sure the original stuff is there as well.

[Fenella:] Great, thank you. "A key to the model is its input which is the result of the LLM splitting into main points. Have you evaluated the splitting model and do you use any tuning of the splitting model based on the survey?"

[Lukas:] If the question asks if we do fine-tuning as such, we don't really do fine-tuning. That's mostly because of our experiences with fine-tuning as usually not the thing you want to calibrate - rather you want to calibrate the context that goes into the model. So that's something we calibrate. Yes so the main considerations is that you want the splitting to, you want each split response to maintain enough of context of original response so that you don't have to read them together. we have made a couple evaluations. Depending on the survey you might not want to split up, as well. You want to group them together. it's a thing that you want to use some judgment on whether you want to use or not. I think that answers the question okay.

[Fenella:] And the last one that we've got so far. is "have any automated methods been used to anonymise participants responses? I'm particularly curious about how Scribe handles personal or identifiable information within the comments."

[Lukas:] Yeah it's a great question and suggestion. We have anonymised them. Part of it is quite simple, of course. If there's any survey parts in the metadata in the survey responses that directly reveal information, we can filter that out...

[Fenella:] Is that still having to be done by a researcher as opposed to the AI model? For example if you've only got three black women in your organisation and you can see that, is that still a researcher having to make that intervention?

[Lukas:] In this case, yes. We haven't implemented any automatic suggestions of that kind. But that's why it's an interesting suggestion, to be honest. I mean, some of that could probably be done. In our case it hasn't been a necessity to because it's an in-house tool and there's always an expectation that we're all going to look through it all. But a lot of these things are possible. And then of course because the responses are split up in that way, there is a further anonymisation, because even the original text is kept hidden if we choose it to be. But one potential interesting thing that can be done is doing a filtering procedure of going through the free-form text and then actually identifying whether there is personal or identifiable information and that's certainly possible. That's not something we've done, but it would be very easy to implement.

[Fenella:] Okay just because we've got a final two minutes I was going to ask a final question which is "if somebody wants to make use of this tool, how do they go about being able to use it? Is it basically if you're a partner or if you're a client of Autonomy or is this something that you guys are looking to commercialise and sell in some form?"

[Lukas:] Yeah, so far we are actually considering making it into a public facing tool, but so far what we've done is we've talked to ... we work a lot with trade unions, for example, and they have a lot of survey responses. We offer this as a service where we get given a survey response or we come in earlier where we actually help, we look at the survey structure and try to shape it to fit the Scribe methodology better and then we do some part of the analysis like the automated stuff and then we work together and process that data. for now, how people use it is to contact us. I can send my email in the chat and then we can start up a conversation about how we can use it together with people. Down the line we might actually try to create some kind of public facing tool / user interfacing tool- so you go on a website and upload a thing. But we're not quite there yet. The demo that we posted is a demo that roughly shows what that could look like, but it's also very rough in its shape. Yeah.

[Fenella:] Well thank you very much for your time. I put a link in the chat to our upcoming seminars and I think the next one is on Social Network Analysis on the 11th of June. That's 2:00. And thank you very much to Sean and Lukas from Autonomy, and we will see you next time. Thank you so much - bye everybody.