Transcript: In Conversation: Mark Elliot and Patrick Sturgis – AI and Social Science

[0:00:00]

Mark Elliot: Hello. Welcome to the latest edition of NCRM's In Conversation series. This is the third NCRM special In Conversation series on AI and its impact on social science. I am Mark Elliot, Professor of Data Science at the University of Manchester, and one of NCRM's co-investigators.

I'm joined today by one of the key figures in UK social science, Patrick Sturgis. Patrick is currently Professor of Quantitative Social Science at LSE and had previous roles, he was himself NCRM's Director and President of the European Social Research Association. He chaired the inquiry into the failure of the 2015 UK election polls and has served as specialist advisor to the House of Lords Select Committee on political polling and digital media. Welcome, Patrick.

Patrick Sturgis: Thanks, Mark. Good to see you.

Mark Elliot: And okay. So, we're going to be talking about AI, which is obviously a hot topic at the moment, and is also a term that gets used and abused and various people mean different things by it. So, perhaps for the viewers, what will you be meaning when you're talking about AI?

Patrick Sturgis: Well, I guess, I mean maybe I'll sort of start off by saying a little bit about what sort of a social scientist am I and what's my interest in AI? I guess I'm probably a bit like you, when we were doing our training, graduate training and we weren't thinking in terms of data science, those terms weren't really around then, the influence of computer science in social science was quite minimal and we would, you know, the sort of analysing rectangular data sets that were nice and clean and so on and often using drop down menus and so on. So, I'm kind of a quantitative social scientist rather than a data scientist.

So just to emphasise that I'm not coming into this, the stuff that we're going to talk about, some of the research that I'm doing as someone with a deep background in data science, computer science and so on, very much more a kind of conventional quantitative social scientist. And I think I've been, you and I have talked, I think, in the past actually in this kind of context where we've been a bit devil's advocating on either side and I've been quite sceptical about big data and so on. I'm more of a kind of a survey person, planned research. So, that's kind of where I've positioned myself often, but I think I'm definitely much more of an optimist or a believer that this is going to be quite transformational for social science, I think, that this is really going to be a step change, as the cliché goes. I think we're already seeing quite a lot of signs of how it's going to change the way that that social scientists work and the sorts of questions that they can answer.

And I guess that also positions me on the more kind of, I'm much more interested in what these tools can do than in I guess the other very important angle that social scientists come from, which is what are the risks and dangers, what are the ethical sides of it and so on. That's not to say I'm not interested in those kinds of things at all, but I'm very much seeing these as kind of that we can build new tools for answering the questions that we're interested in in social science.

And that of course, you know, AI has been around for a long time and as you say, it means different things. So, I'm really talking here about large language models and the huge number and quantity of these things that are now emerging.

Mark Elliot: Yeah. So, and I think, I mean, it's always worth making this point that because that has been kind of forced into public consciousness, large language models and ChatGPT in particular, that then becomes synonymous with what AI means to many people. And in reality there's a much wider set of things which we mean by AI.

I was just reflecting on your comments about the more traditional approach to. I of course originally came from an AI background and that meant something very different in the 1980s when I was doing my research and actually getting time on computers to do AI research was limited at that point and we had to do a lot of stuff on pen and paper and working out what algorithms would mean and so forth before we even got near a computer. So, I think it’s worth reflecting that this was the initial steps of AI were very, very much theoretical rather than you've got this kind of image of AI researchers sitting in their white coats next to a humming computer. And that just wasn’t the case.

[0:05:39]

Yeah, okay. So, I mean, in terms of your own adoption of AI, in terms of your workflow, how's that been?

Patrick Sturgis: Yeah, well, I think, I mean, that's an interesting point you make there and I'm aware of your kind of very longstanding interest in and work on AI. And as you said, I think now we're already at the stage where social scientists, not just people in humanities, any disciplinary area, you can approach these as tools that you can use without really needing to know very much about what's going on under the hood. You know, I mean, that's all very interesting, so I mean that's computer science, right? But we can take these tools now and use them, deploy them, very flexibly and I mean that's, I guess, some of the things that we'll talk about now, the sorts of things how I'm using it.

I guess, so the first thing I think I want to talk about would be, yes, the pipeline, the workflow that now is accessible to someone like me and going back to my earlier comments about I'm not someone who codes freely in Python. I’ve had this difficult journey of learning SPSS. I was very, very proficient in writing SPSS code and then that becomes very unfashionable and I moved over to using Stata and I learned how to be very good at Stata and so on. And then R comes along and I'm kind of like, “Oh, bloody hell, I'm going to have to learn something else now. I'm getting too old for this”.

So, I was never really great at coding in R and certainly not in Python, but what I've found is that with the help of ChatGPT, or other large language models are also equally good or even better at this, are what's called vibe coding in the lingo. I don't think that's the greatest name for it. I see where it comes from, but I think I prefer to think about it as kind of goal oriented coding that you're, because I think that's important and you'll see, I mean I think you know what I'm talking about, but in order to do this, it’s not open to anyone because you need to know what you're trying to do, okay, you need to know where you're trying to get the thing that you're trying to, if you want to analyse data or you want to build a particular kind of chart or you want to set up some pipeline to scrape text from documents and so on. You need to know what the thing that you're producing should look like, but if you do that now through just language coding, just through telling ChatGPT what you want, it will take you through the process. It will, you know, and I did this a little while ago.

You know, it told me to install Anaconda. It's told me to open up Jupyter Notebooks and it told me to put this code in here and run this and sometimes it wouldn't work and you'd put your error message in. I mean it’s really incredible how it can take you from not really having those skills that people toiled to develop, that you can now access this kind of functionality.

And so, one of the things that I'm now able to do is, quite quickly as well, it’s a very quick process that you can build these things is that interacting with large language models through an API rather than typing directly into the query function and getting individual responses that you can massively scale things up, particularly sort of things like sending documents, extracting information from them, analysing them in various ways, outputting it in some other format and analysing it.

So, I think that is a really, really exciting development. I mean there are obviously some risks and problematic parts of that which we wonder how much do we need to know, be able to look at that code and understand what's going on. That's an open question, I think. But largely you can judge it by how successful it is. If the thing that it produces, if it does end up building the pipeline that you want in a way that works and produces the outcomes, I think who really cares if it's not as neat or efficient as it might be?

And so, yeah, that I think opens up new ways of working for people who otherwise would have been limited by the need to be constantly upskilling with whichever is the latest new bit of code that you need to learn.

[0:10:43]

So that is, as I say, that’s one thing. The coding and what that's going to open up to social scientists, I think, and I think it's probably quite limited at the moment, the sort of people who are experimenting with that probably still seems a little bit way off for a lot of people, but really it isn't. And I think, soon enough that's going to be quite widespread and that will really as I say, scale up the sorts of things that we can do.

Mark Elliot: I think what you said earlier on about having some understanding of voice kind of thing, it's analogous to when the calculators were introduced. So, I reveal my age because this happened whilst I was at school and there was a big kind of, “Oh, is this a good idea?” kind of thing. And I've noticed, as time has gone on in, in the students that I'm sort of working with, even PhD students, they're not necessarily able to understand that the number that they've got out of a calculation is not correct just because it looks wrong, which I just instinctively know that can't be right, something must be wrong. There's this just automatic belief in it because they haven't gone through that stage of doing mental arithmetic which is almost a kind of artefact of a bygone age now, and it's in this kind of category we're in now because you're better able to use LLMs to do coding if you can already code than if you can't because you kind of understand what the structure should look like.

But I do think that we're getting to a stage where it's more like driving a car than it is being a car mechanic, whereas before you were definitely a mechanic if you were coding, and now you've got to know how to get it to go in the right direction.

And I think it will get smoother and smoother as time goes on and as more people adopt it as part of their practice.

Patrick Sturgis: Indeed. Yeah, I mean we don't look at the C+ underlying the Python code, do we? I mean so it’s just a development, the next layer of, and we’re now at a point where that top layer is just telling it what to do. As I say, that isn't enough on its own. You do need, I mean, I think you still need to, if you want to run statistical models, you need to understand what the model is and what the data structure is and you're in real danger of just producing some junk if you don't know that. But just the mechanics of putting the code together is now much more accessible to social scientists and I’m looking forward to seeing how that. I mean, it has some kind of other interesting knock-ons which are like what do we teach students? Here at LSE in the methodology department, we have courses on basically teaching people how to code in Python and in R and so on and we're already wondering how long are we going to be teaching this stuff? We're going to be teaching something along those lines, but it's not going to be here's how to write a loop and you know? So, there’s some interesting unintended or hard to predict outcomes as well.

But maybe I can give you an actual example of some of the research that this has opened up to me, and this is something I've been working on with Ian Brunton-Smith at the University of Surrey. Ian's a longstanding collaborator of mine. And actually, when I was, as you mentioned, I was at NCRM when I was at Southampton and part of that research, we wrote a paper, myself and Rebekah Luff, and this was based on coding the methodological content of journal articles. So, it was building on some previous publications where researchers had done this for particular years back, going back to the 1950s and then updated in the 1980s and 2000, looking at the proportion of published articles that are theory compared to empirical and of empirical what are qual and quants and using surveys? So, we particularly looked at surveys and there was, I think 1,500 journal articles downloaded by hand and then hand coded by PhD students, and we had a team of seven PhD students coding these things and we did a little code of reliability study. We’re sort of glad we did that. And yeah, we wrote an article and it's kind of interesting and we got some citations for it, surveys still going strong.

And anyway, when these large language models come along and I'm thinking, “Yeah, I'm sure these seem like they could read a journal article and tell you whether it's theory or empirical”. So, we built a pipeline using Jupyter and Python and scraping the content of the articles, going into its credit, and sure enough we've got the human codes and we can line them up and see that the LLM does as well as a human, or indeed better in many cases.

[0:16:19]

So, and actually, as you do in these kinds of situations where you've got no truth value, if the human says it's an online survey and the LLM says it's a postal survey, the only way to kind of reconcile those is to go and read it. So, Ian and I went in, went looking at some of these discordant cases, went and had a look, and actually oftentimes it was the LLM that was wrong, but nearly as often it was the human that was wrong and that the LLM was getting the right answer and that kind of makes sense, you know, because if you're the human, you're going to get bored, you're going to get tired, you've got an incentive to be quick because you're going to get more money the quicker you get through them and LLMs don't have those incentives.

So, we were able to code the full set of 1,500 articles and this had taken the humans, I think it, well, it took over three months to get it done because of course the humans can't do it all in one go. They need to take a break and go through. And LLM did this in under ten hours and in effect it was free. It cost NCRM I think £10,000 to pay the researchers. The LLM did it for free in ten hours.

And of course, what we were then able to do is having demonstrated that this works as well or better than a human, we can then scale that up almost free. Of course, you start paying for tokens and so on and depending on what kind of a model you want to use, maybe a reasoning model’s a bit more accurate, a bit slower and so on. But then you can code tens of thousands of journal articles and do it extremely quickly, analyse the data.

So that's a project I've been working on which is using this sort of vibe coding and the kind of text classification abilities of LLMs.

Mark Elliot: Yeah, I mean I think that's sort of what I kind of loosely call qualitative analytics at scale. I think it’s one of the things that gets opened up by these new approaches. It's always been an issue for qualy researchers that it's difficult to get sufficient quantity of data to make generalisable conclusions and that's kind of been the constraint I've been working with. But this is a set of tools which now open up that possibility with the caveats on that, obviously, and I suspect if we had a qualitative researcher in the room, they would have something to say about that. But it certainly presents new opportunities that just simply weren't there before. Yeah, yeah.

Patrick Sturgis: Yeah. And as I say, once you figure out how to interact with an LLM through the API so that you can just send up very, very quick repeated requests and dump down the responses into a CSV file or something and analyse that, again it really transforms the sorts of things that you can do and that you can even think about doing, but you wouldn't even think about it because there's no way you could do that.

So that's one area. Another area where I have actually an ESRC grant with my colleague here, Tom Robinson and Caroline Roberts at University of Lausanne, and that's part of the Survey Futures programme. And what we're doing there is looking at how generative AI can be integrated into survey research, into some survey workflow. Obviously there's a huge amount of interest in the world of survey methodology about what LLMs are going to be able to do. In fact, I’m going out to St Louis in a couple of weeks for the American Association of Public Opinion Research AAPOR conference. I go there most years. I think last year there may have been four or five papers on LLMs. Looking at the programme, I'd say that's probably about 30% or 40% this year. It’s almost a conference on LLMs in survey research. So that's just as an indicator of I think the kind of excitement about the potential here.

We're looking at a relatively narrow part of this which we think will be, you know, where LLMs are very well suited. And that's as kind of interviewers or indeed as respondents.

So, we've got a few bits of work packages in the jargon, the first of which I'll tell you a little bit about, which is coding occupation in surveys. There's a notoriously challenging thing to measure in a survey because what you're trying to ultimately measure, as you'll know this, but the Office for National Statistics or any other NSI will have a very long list of all the occupations that there are in an economy and this will run into hundreds or even thousands of different occupations. And they're all described, not all, but many of them are described in quite technical ways which would be familiar with the person who does that occupation.

[0:21:55]

Okay, so these two features, the very, very large number of them and the technical way they're described means that we can't do the usual thing, which is here's a list, pick your one. Or we can't even do that very well with the dynamic list where you type in or something like that. It just doesn't work. So, the way that we've always measured occupation is to ask people to give open responses about their job and what they do in their job and the industry and so on. And so, this is quite inefficient because it means that you have to ask every respondent a whole bunch of questions and for many of them you wouldn't need all of those questions to do the coding so that the interviewer or, in the self-completion survey, the respondent gives these answers and they're coded at a separate stage in the office and there's a team of humans whose job is to find the right kind of four digit code given these responses. You know, labour intensive slow.

So, what we're doing is thinking, well, we could use an LLM here and integrate the measurement and the coding on the fly. So, we've built a pipeline that has the LLM as part of the questionnaire scripting software. So, if you're my respondent, the first question would be, “What's your job title?” You say professor. That response then gets passed to the LLM. I won't go into the full detail. There's a RAG model in there as well. But in effect, the LLM tries to code it to the 400 listed. And we give it some information about how to do this in the prompt. It tries to find the right one.

Often it can do that. It’s pretty accurate, it can find the right code. So that's it. You don't need to answer any more questions. But if it can't or it can't do that with confidence, then it comes back and it's going to probe you, just like an interviewer would, but it's going to probe you to say, “Okay, so are you a professor in a university or in a research institute or something like that?” Something which would enable it to break the tie between multiple equally plausible codes.

And so, this has several benefits. As I said, one is it's going to reduce respondent burden. These are particularly burdensome questions. Respondents don't enjoy putting them in because they have to type the open responses. They often do that poorly, they don't give very much detail. So, it will reduce burden. We get better accuracy with the codings. What we're looking at thus far suggests that they're going to be more accurate. But also, you can then target the follow up questions so that you can, and that might be sometimes because the person has given just a bad answer, they've typed in a typo in there. So, it could come back and say, “Could you say that again or can you give a bit of detail about what your company does?” that sort of thing.

So, and we've built this, we're going to be kind of like trying this out on real humans in a survey hopefully towards the end of the month and we're going to get some data and write a paper up on that.

So, I mean, we're aware a number of people, a number of NSIs are working on this, but that it has that basic integration of the LLM into the questionnaire opens up lots of interesting possibilities because you now can start having kind of dynamic self-completion questionnaires. That's one of, of course, the limitations of moving from an interview administer to a self-completion survey. So, you don't have the interviewer there to explain, follow up, asking probing question tailored to the particular respondent. So, we can maybe now start doing kind of more conversational interviewing at scale.

What we're specifically interested in in our Survey Futures project is can it be used for cognitive interviewing the way that we can use this cognitive interviewing technique to evaluate draft questions. That's very widely used in the industry. It involves getting respondents to think out loud when they're answering the questions. So, we'll present them with a draft question, say, “What’s going through your mind when you answer?” They say, “Okay, well I’m a bit confused by what this word means here”. And okay, that's a red flag, and if that comes up in several interviews, then you see it.

[0:26:33]

So, we've got a tool that a colleague of ours in the methodology department here at LSE has built, Friedrich Geiecke has built a kind of an LLM robot that will interview people qualitatively. So, we're going to fine tune this to make it focus on the sorts of ways that cognitive interviewers ask questions and get this data and we'll see how it works. I mean, I think there's reasons to be sceptical about how well LLMs can act or can replace humans. But we'll see.

And then where it gets a little bit weird is then we're also going to have other LLM robots being respondents. So, we'll prompt it to be a particular characteristic of person, say you're someone who doesn't have a degree, who lives in the north of England, is single and so on, and we get different kind of data from that person than someone we prompted to be a different kind of demographic.

Then obviously one of the things that this enables is this scaling up again. If we can have an LLM as the interviewer interviewing multiple LLMs as the respondent, then the usual kind of restrictions that we put on cognitive interviewing, I don't know, you maybe do 20 interviews, don't do all the questions, those aren't because we think that's the best thing to do. It's because it's too time-consuming and costly to do more, particularly if you've got cross national surveys. We just do one or two countries.

Now again, these machines can translate on the fly as well, so we could really scale up what we can do in terms of question testing. And I’ll just say one more thing about it is that I think people might object to this, I think, reasonably enough. Can we really replace human respondents, human interviewers? I think my current take on this, and this may change as we get more into the research, is that probably not. Probably humans are going to be better, but at the moment many people who, you know, students or people with low budgets, research students and so on, can't afford to do it at all. So, this might be something that would be available that would give you a good take on where the problems are in a draft questionnaire and isn't necessarily the gold standard but is very scalable and in effect free.

Mark Elliot: Yeah, absolutely. And I mean, that's one of the things that gets us around back to almost the driving analogy where you’ve still got to keep your eye on the road, but yeah, absolutely you can use this to help you with drafts and to kind of move you in a better space and then you go to the humans and then that's that premium time that you're now using to kind of really refine it. So, you could hopefully end up with something that is better than if you do either of the things but not both of them.

I did actually, just touching on something which is of interest to me with a slightly different hat on, which is about augmentation. And you can think about the possibility that genuine respondents to surveys might actually start to use AI to help them to respond to surveys. I can't remember what my income is, but my AI assistant knows and so you can kind of think, well, at what point is that something different that's actually responding to your survey and what is a data unit in the context of AI augmented humanity?

And I’m going to just finish off now with a topic which I think you probably also have some sort of side interest in, I would imagine, given your focus on public opinion, is often in the news these days as well as the impact of AI on public opinion, affecting elections and something. And do you have a view on the way that’s going?

Patrick Sturgis: Do you mean, well, there's different ways you can interpret that. I suppose one is in coming back to the point you made a moment ago about people using AI LLMs to respond to surveys. I think this is a real challenge for the survey industry. There's always been bots that people use, but whenever you're giving financial incentives for people to complete surveys or do any task, there's always an incentive for the people to automate that and get the incentive without the actual work of being a human. And so, the whole industry that's based on people opting in, you know, I'll do surveys for you, I'll sign up to your panel, I'll click on this thing and I'll get some benefit from it, there's a huge, huge challenge how they're going to prevent LLMs basically cannibalising their panels and so on and there’s already a very, very large concern.

[0:31:45]

I have colleagues who are, this is for another talk, but getting humans to interact with bots and bots with each other and seeing how they all get on and so on. But a lot of them, when they're getting their bots to interact with humans on these panels, what they're actually concerned is these aren't humans anyway. They're actually someone else's LLM here is the person who.

So, I mean and there's another interesting consequence of that is that we might see a bit more of a resurgence in the kind of random sampling approach, not so much because of the representativeness, although that's an important part, but because you can build a wall around who your respondents are, you can't just opt into it, you have to be sampled. I've never been sampled for a random probability sample in my working life as a survey researcher. So, it's a very low probability event getting sampled for one these. So that’s another kind of, you know, how methodologies might interact in the future.

But I guess I think maybe your question was more about misinformation and that sort of thing in terms of influencing people around election time. It’s an interesting question. It's not one that I'm kind of actively researching. There's a huge amount of interest in this topic of detecting information, misinformation and so on online, how that affects opinion, a big source of concern, but not one that I'm particularly focused on myself at the moment. I'll let others worry about that.

Mark Elliot: Yes. And just as an amusing anecdote in response to your never being sampled, I actually got sampled two years consecutively for the labour force survey at different addresses, which is about equivalent to winning the lottery I think in terms of probability.

Patrick Sturgis: Yes, indeed.

Mark Elliot: I've used up all my lottery luck in my lifetime. Okay, well, thank you very much, Patrick. That was a really interesting conversation and I think our viewers will be also very interested in it. So, I'll say goodbye for now.

Patrick Sturgis: Thank you very much, a real pleasure to talk to you.

[End of Transcript]