Transcript: In Conversation: Mark Elliot and Alex Singleton – AI and Social Science

[0:00:00]

Mark Elliot: Hello. This is the fourth in NCRM's In Conversation series on AI and social science.

I am Mark Elliot, Professor of Data Science at the University of Manchester and co-director of NCRM.

I'm joined today by Alex Singleton, who is Professor of Geographic Information Science at the University of Liverpool and where he is Director of the recently launched Geographic Data Service. Alex's research has straddled the boundary between social and computational sciences and so he's a key person for this series. Alex has, for example, extended the science of area classification within urban analytics examining how geodemographics using unsupervised machine learning can be refined for effective yet ethical use and how systems comprising AI can assist in public resource allocation applications. Welcome, Alex.

Alex Singleton: Thanks, Mark. It's great to be invited along. So over to you for some questions.

Mark Elliot: Okay. Well, I'll start with the question that I ask all guests on this series. AI is a big topic area and it's always a term that gets used by different people to mean different things and also abused. So, for the viewers, what will you be meaning when you talk about AI?

Alex Singleton: I think in its broadest sense, my interpretation of AI is any type of computer code that's aiming to sort of mimic human behaviour or human like behaviour, sort of perceptions or understanding. So, it has some degree of understanding, interpretation and then answering of a question.

Mark Elliot: So, you use the word “mimic” there. So, it's not necessarily the case that we are assuming that AI is actually intelligent, it's just that it behaves as if it is intelligent.

Alex Singleton: Yeah, I'm not buying into conspiracy theories about AI being sentient just yet, but so I still, I use the word “mimic” very specifically for that reason. So, I think we're quite a long way from the kind of view that this is a clever autocomplete and it certainly has a lot of very useful applications, but I don't think we're close yet to artificial general intelligence, which I think is a slightly different thing. That still feels like science fiction. But to be fair, a lot of things that we can do now with AI and new AI models probably would have felt like science fiction five or six years ago as well.

Mark Elliot: Yeah, absolutely. I think the difference that I've noticed is how people are talking about it and interacting with it. So, what's happened of course in the last five years is now everybody who has a smartphone or some other device has AI on it, is interacting with AI. So, what has been your personal experience of using AI?

Alex Singleton: I mean as it transpired for a long time without thinking about it, a lot of my research was AI. I mean you mentioned sort of I've had interest in building geodemographic classifications which are principally created by unsupervised machine learning, clustering algorithms. You know, they broadly fit within the sort of AI landscape.

But my background is, I did a geography degree and I did a PhD with a fairly traditional sort of GI science group. I had some exposure to computer science as part of that and I was sort of vaguely aware in my discipline things that had sort of happened in the late 90s and early 2000s, this area called geo computation that in different parts had sort of dabbled in AI as well. So, I'd familiarity with these things, but I wouldn't say initially when I started it was really core to what I was doing. I've only reflected on this after the fact.

I think when I moved to Liverpool to get my first academic position here, I was interested in developing a brand for Liverpool around the things that I was starting to get interested in. It was also at a time when data science was a new interdisciplinary area and it was kind of growing. There was lots of data science MScs popping up across the UK and also globally.

[0:04:39]

So, with that in mind, I set up something here called the Geographic Data Science Lab and that was really trying to ensure that geographers weren't kind of left out of this new economy. And we set up the first programme in that and then also I built a quite a large group around that particular topic and we recruited academics to it.

So, I got more interested in the computational side of what I was doing perhaps than what I'd done previously after I became a lecturer here. And I started working collaboratively with people in our computer science department on a couple of projects. And that sort of got me started on AI properly.

So, one of the first things that we did, we got some funding, I forget where from, to look at Street View advertisements. So, we built some convolutional neural network models to try and pick out from Street View imagery where the locations of fast food advertisements were. And then from that we can actually do things like measure the proximity of those to schools because there's some legislation around planning that means you shouldn't put fast food advertisements near schools and there's some considerations and constraints on that. And that was a really nice project because it had a sort of computational element which pleased the computer scientists. I've learned some more stuff about AI at that points, but it also linked to some of the core things that geographers like to think about and also a sort of social and an applied angle.

And actually, as part of that we also did some really, we had a very good postdoc working on this, incidentally, did some very, very interesting stuff which actually got me thinking a little bit about the relationship between AI and data in that one of the one of the problems is we needed data to train these models that we were doing, but they didn't exist. So, we actually created ways of building synthetic data and actually created a synthetic training data set where we would take existing food advertisements, we would take existing pictures of billboards, and we would essentially fuse those two things together in a model which then became training data to use in our predictive model to identify where these locations were.

So, I learned a lot in that project about what was possible. And this is many, many years before we knew what ChatGPT was and generative AI. And that was really the first sort of thing that we did.

The next project we worked on is perhaps a little bit more closely tied to what geographers might be interested in in that we took satellite imagery and had a new look at the way in which you might build land use classifications. So, this is if you take images from space, can you work out what the patterns are on the ground? And actually, if you look at most land cover classifications or land use classifications, there's quite a strong degree of focus on things that aren't the built environment. So, you might have a differentiation between urban and suburban. I mean that's not universal and there are exceptions to that rule.

So, in that project we looked at the problem slightly differently. So actually, what we did was we, rather than taking every pixel which covers the whole of the UK, we just looked at those areas around people's properties. So, we did it actually at postcode level and we took a sort of square grid around every single postcode in the country. We took the data for those areas which was a red, green, blue and near infrared band, and then we used something called a convolutional neural network again to essentially create a representation of that block of data. And then from that we use that into a clustering model like we would in a traditional geodemographic and we built this focused classification of postcodes purely on satellite data. And it's actually quite different from how people who build land use and land cover classifications do their models. And again, it was just that was another kind of quite an interesting application of AI.

You'll note the techniques we're using are similar, but we're using them in quite different domains and actually we've got, you know, I think the original idea for the image classification from satellite data came from another project which my colleague, Danny Arribas-Bel, and I had with a PhD student who was actually looking at using convolutional neural networks to look at images of shop frontages and we were seeing whether we could plug those into hedonic models to predict rental prices for properties.

[0:09:29]

So, these new techniques and toys I've just found incredibly interesting to bring into my own research and adapt. And perhaps I suppose if I've got time to talk about one more example, one of the most recent things I’m interested in at the moment is how you can bring together multimodal data to say build a new geodemographic. So, I've been working with Stef De Sabbata at the University of Leicester on a project where we're using, at the moment, using auto encoders to bring together census data and to try and pick out some of those interesting nonlinear patterns that you get in geographic data because of the nature of geography, and then bring those into a geodemographic classification to try and improve geodemographic classifications representation of what the world looks like.

So, an early testing of all this is it's very, very effective. So, one of the problems that you have in geodemographics when you're using traditional methods is you tend to find if you're building a national classification and you don't treat places like London as an example differently because London is different from the rest of the UK, it doesn't actually look very good and doesn't work particularly well in London because essentially the London averages for all the variables that you might put into a geodemographic are very different from the other parts of the country for the same variables so it has peculiar effects. And actually, some of these new techniques that we're using auto encoders for get around some of that problem because they're actually much better at understanding those nonlinear relationships between the variables across space and as a result the end classifications both look better and actually perform better.

So, we're in the process of writing that up at the moment, both in terms of building these models of the structure of place, but then also following through and using those in geodemographic classifications as well.

Mark Elliot: How fascinating, Alex. I was very struck as you were talking there that a lot of what you're describing is essentially augmentation of your own capacity and increasing the sort of research productivity and your capacity to do things perhaps that you couldn't do before because of these new tools. I mean that definitely feels like AI is a tool that you're adding to your toolbox. It kind of seems akin to the discovery, invention of multi-level modelling in the 1990s I think that was, kind of changed the way we thought about doing modelling. And some of the descriptions you've got here are akin to that, we’ve got this new way of looking at that which suddenly enables us to do different things with data that we weren't previously able to.

So, your service, geographic data service, I’m very interested to know a bit more about that and how using AI in that.

Alex Singleton: Yes. So, I mean, for the past, I'm trying to think, probably 10 or 11 years, I ran something called the Consumer Data Research Centre, CDRC, maybe a number of the viewers are familiar with that because we held lots of data that were neither administrative nor survey, traditional survey based data. So, they sat within the middle. So, it would be things like consumer data or commercial surveys that weren't government surveys and mobility data, different types of image data. We were funded under something called BDN2, Big Data Network 2, which they ran over that period of time.

And then more recently that transitioned into an opportunity through something called Smart Data Research UK and the same team who ran CDRC at UCL and Liverpool, got the same team together and we put a new bid in to continue some of those services but also extend them into some of the new areas off smart data.

So, we formally launched this in April with a brand new data catalogue and we've got a maps platform. And essentially the service continues in a similar vein to CDRC in that we've got data that's open, safeguarded and secure and we have a whole series of governance and protocols around how you can get access to that data. And then we’re continuing our mission over the next couple of years and particularly around the focus around online data and image data along with some of these other traditional data sets that we've got using geography to integrate these together.

[0:14:26]

So, what we're quite interested in doing in the geographic data service is bringing together lots of these complicated data sets using innovative methods and linkage methods to produce things that will be used by a lot of people and will be valuable. So, if people have got ideas for things that they would be interested in, data products that maybe don't exist, definitely get in touch, we'd love to hear about that because we're always trying to produce data products that people actually will want to use.

And in terms of AI, I mean most of the work that we're doing now is infused by AI methods in different ways, just because of the way in which we actually do social science now.

But perhaps something that’s more explicitly using AI, which I think is really exciting, is we were fortunate to win, and this is my colleague Mark Green here at the University of Liverpool, some funding from ESRC’s Future Data Services to look specifically at the idea of semantic search of data catalogues. So, we've got, I don't know, let's say a hundred data sets in our data catalogue. UKDS have got probably many hundreds of data sets, and then there's a number of others, ADR UK the same. So, there's all these data catalogues out there and they're not connected together. And what we actually have done as part of our geographic data service is use something called an embedding model to enable semantic search.

Now there's lots of fancy words there, but an embedding essentially is that you can send a block of text, and in this case a description of a data set, to an embedding model and the embedding model turns that block of text into what will be best described as sort of a magic number. And that magic number represents what is in that text record. Now what you have to do is you have to do that for every tax record that you've got. So, you have a semantic embedding for every single record in your data catalogue. And then when somebody searches for a particular term on the data catalogue, it turns that into one of these magic numbers, and then you can compare these numbers and actually produce a relationship between the search query and the embeddings that you've got related to your data records and your data catalogue.

And where this is really clever is it gets you beyond just saying if this word is in, you know, if this search word is in the record, we'll bring that and make that record relevant. It actually looks for semantic similarity. So, a very good example of this would be one of our early tests was around the term “diabetes”. Incidentally, we've been through a process of improving our metadata in the Geographic Data Service now, so actually traditional search actually works much better. About a year ago, it wasn't so good. So, we would say search for something like diabetes, zero results would have come up on our original data catalogue because we had no metadata in there which had the word “diabetes”. In the semantic version of this, when we put that same word in, with exactly the same data just with these embeddings attached, a whole series of different records show up which are related to that term. So, it might be survey data about diet and food consumption, which although it doesn't have the word “diabetes” in, there is a semantic understanding that often people who have diabetes, that may be an influencing factor. So actually, if you were wanting to research that, this probably is a relevant data set.

So, we're using it quite explicitly in that context. And actually, we haven't ported that over to the Geographic Data Service yet, but that's one of the things that we will do fairly shortly now that the main data catalogue has been launched.

But with that project in particular, one of the things we're also interested in doing is bringing together multiple data catalogues so you're not siloing things, because we're interested in looking across UKRI’s data landscape. So, there might be data in NERC, there might be data that's from the MRC and there might be stuff that people have done previously in UKDS and you might want to bring all of that together actually to be relevant to your particular question. So, it's a more natural way of actually searching for data that doesn't just necessarily rely on keywords alone.

[0:18:42]

Mark Elliot: Yeah, absolutely vital sort of work. And because I'm involved in the Future Data Service programme myself, I've kind of been keeping a watching brief on that. It's come up again and again in the data resources training network as an issue, and you can imagine a PhD student on a journey trying to work out what data they're going to use for their new project, it takes a really long time and it's very manually intensive and not particularly a critical part of the research process. Well, it is critical, but it's not the kind of inventive, creative part of the research process, so finding ways to improve that is absolutely critical. So, it’s really interesting that that project has pursued that avenue because I think there'll be a lot of use of that sort of technology.

Alex Singleton: Yeah. And I mean, I think it's also slightly, I mean, it's reflective of what's going on in general. I mean keyword search is not going to be a thing in ten years’ time. You know, search engines are going to be AI engines and this is how essentially we're going to consolidate and feed information forward. So, I sort of feel like we need to be ahead of this because, you know, when I was an undergraduate, I remember my undergraduate tutor, this is in the, I mean, you know, I'm old enough to remember the time before search engines, would show me how to search a search engine, which is, you know, we all learned at that time the way you write a query to get a good result back. And this doesn't need that. You can write a query for these things and however abstract or scrappy as you like, and it will understand what you mean and the intention of that query, or try to understand the intention of that query and it will find things that are relevant to that and actually represent it back, maybe synthesise materials for you.

So, I think this project that we've been working on really is sort of moving in the same direction in which information consumption is happening more generally in the kind of widening economy. So, I do think they are quite important. But you know, like all things of academia, there is pushback from people who like the traditional way of doing these things through things like knowledge graphs and meta. Actually, the way you solve this problem is better metadata.

We've got to be careful we don't fall into that trap of old ways of doing things because actually this is such a fast moving area in industry, we'll just be left behind and look like Luddites. So, I'm very keen that we actually get this working and implemented and promoted as a way of doing this stuff because it solves lots of problems that people historically have spent a long time in. The actual necessity to have extremely standardised and structured metadata is lessened in this new model which actually for some people who've based their careers around these things I think is a controversial topic.

So anyway, we'll see how things go, but we're very keen to have this functionality as part of the Geographic Data Service going forward.

Mark Elliot: I think it reflects a bigger picture, as you say, in the sense of this kind of move away from static ways of describing and capturing things. Interestingly, we've just had a big discussion in another one of those networks with DTP training network about how to give advice to PhD students on AI. And after a lot of discussion, we decided we weren't going to produce any static resources for that because there just wasn't any point, you know, they'd be out of date as soon as you've written it and it doesn't really kind of fit to what a student needs, which is a way of kind of helping them to think about how they're going to use AI. So, it's definitely a much more dynamic resource we're going for in that place. So, it’s like a kind of spillover effect into thinking about meta resources as well as the sort of stuff you're talking about.

It’s really interesting and that relates to another thing, I think, which you've kind of drawn in is changing the way we think about information and data because they too are not static things. We're used to kind of 20 years ago, perhaps, you look at this model, you kind of got a data set and then that was your data set and then you analyse it to death. And this is probably not what we're going to be doing in in five, ten years’ time. It will be much more of a dynamic interaction.

[0:23:06]

You mentioned synthesis earlier on. Obviously that's another thing which is feeding into this where the data is not just that kind of literal, this data is about something and represents this person or this other population unit. It's a dynamic representation of an environment. So, I think it's kind of really, really interesting the way all these things interleave in the development. And keeping track of it all is obviously horrendous for any individual.

Any other projects you'd like to mention where you've used LLMs or other AI?

Alex Singleton: Yeah, I mean the most recent live one again has been really interesting. So, I mean I've been working with our colleagues here in the planning elements of our geography and planning department where one of the things that if you're a local authority and you're introducing a new planning policy document, so something like a local plan or a supplementary planning document, you have to go out to the public and ask them for their opinions because the idea is that if you're making changes to the structure of the places in which people live, in a democracy you should have some voice that you can kind of raise concerns or just general comments or support these particular plans.

So, there’s a whole sort of statutory framework around this where planning authorities, which can be a local authority, or in our case actually two local authorities together as in Greater Cambridge are planning, have to ask the public for their opinions on these plans. So, there's online tools. The general public can submit responses to these consultations. And in some of the consultations we've been looking at, we've been doing a pilot project around three consultations in

Cambridgeshire. You know there's been sort of a, you know, into the sort of multiple hundreds of responses to them. But actually, for very big local plans, my colleague was saying, I think it was in Greater Manchester, I think it was some large planning consultation they did there where there was I think 3,500 responses. So that's a lot of text. This is principally text, attachments to emails, emails, summaries of larger blocks of text. There's lots of different things. And legally, the planning officers within that planning constituency have to go and look at all of these and they have to do two things.

One, they have to synthesise each individual’s response which could be across multiple documents. And then they also have to take those and then create a summary report of the general feelings, queries, complaints, suggestions from the public aligned to the planning documents. And you can imagine that in some of the work we were doing where there was sort of a couple of hundred, it was about 60 hours of a planning officer's time to do all of this manual summarisation. What we found is we've built an AI tool that can do it in ten minutes, the whole thing sort of end to end. So, you can actually save a huge amount of time which frees up planning officers to go and do more interesting things, because it's quite mundane.

And then the other thing is that the actual quality that you can actually get in terms of the summarisations out of the LLM is much more detailed and actually more concise than actually the human can do. And the reason for that is, I mean, I don't know how you'd respond to having to read hundreds of complaints or comments from the public about something over the day, but you're never going to be able to capture all of that cognitively in your head, whereas a computer can and it can bring all those things together and summarise them quite nicely.

But obviously because there’s a legal angle, a statutory angle to this, we obviously have to be very, very careful about how you do it, because there's all sorts of things around data governance and GDPR and then there's the issue, particularly we’re if using LLMs to do these summarisations of the problems with hallucinations. So, what we've actually done is built a whole pipeline system with all of those considerations baked into them. So, we do lots of things where we're double checking our homework and making sure that actually what is summarised is a true representation.

[0:27:26]

We do all sorts of other different bits and pieces behind the scenes to ensure that the outputs are consistent and reflective of the sort of the tones of what the public is saying. We do different types of sentiment analysis where we can look at whether these are just comments or oppositions or supports. And then also summarise all these into some nice executive summaries and reports that are just downloadable.

So, it's been, I mean it has been a really good project. We made sure that when we were doing this it was fully embedded and in partnership with the Planning Authority. So, they've been working with us on this sort of every step of the way. And then also we have done a number of public consultation workshops around Cambridge as well. And actually, all of those learnings have been very helpful to ensure that we mitigate the public's concerns around the use of these tools because essentially, if these were adopted more widely, they are going to summarise and provide information for the Planning Inspectorate to make decisions on and whether these plans are sound, which means if they're sound, they can be implemented. So, you've got to bring the public along with you and try and listen to their concerns and articulate responses to those concerns and address them in the software that you're building.

So that's been a really interesting project. Again, another slight deviation from what we've done previously, but again really shows the power of what some of these new platforms can actually do if managed effectively.

Mark Elliot: I think that's a whole area of research in itself, thinking about the public’s perception of AI and its relationship with AI, and obviously really important and speaks to an area where social sciences might itself have something to say thinking about that sort of relationship between technology and humans as it's developing.

We need probably to close our session now. Just to finish off, I know you've got some views about the UK's position in respect of AI and social sciences and whether we're putting enough resources into it. Perhaps you could give us some closing thoughts on that.

Alex Singleton: Yeah. I mean, I think reflecting on, you know, so I mean you mentioned that social scientists can have a kind of role in terms of the ethics of AI, and I think that's an incredibly valid thing for us to be doing and actually I think we probably shout loudest as a community about that. But I think where we will run into trouble is if that's the only thing that we do and we're actually not training our students to use these new methods, but use them critically, because if we don't do that, they won't have the skill sets for the new economy and what's to come, and essentially we'll have a reduced relevance in this.

So, I think I'm very keen that we have a critical view, which I think is what social sciences are good for. And I think it's very important that we're giving people the skills that they can implement these types of new techniques safely and ethically.

I do worry a little bit about how this national voice is presented. You know, we have a National Institute for AI and Data Science through the Alan Turing Institute, and increasingly over the years, that has withdrawn support for some of the social science elements which it previously supported. So, there was a very large sort of urban analytics programme. My understanding is most of that now has gone. I think I read last week that some of the work around data science, AI policy, I think that is now much smaller than it used to be as well, and most of the activities seem to be more focused around areas that are perhaps less relevant to the social science or there's fewer social science voices in those kind of audiences.

So, you know, that is the National Centre for Data Science and AI where I don't feel social science is particularly well represented. You know, we do have the Ada Lovelace, which I think is another excellent activity, but again that is more around the ethics of AI. And again, I think we miss some elements of actually social sciences doing these things. Because so, you know, I'm a social scientist by background and have kind of picked these things up as I've gone along and I think we're doing some really interesting research here. And I think that opportunity should be made available to everybody else as well. But I do worry that we will be forgotten about because of the very rapid growth and interest in this area unless we actually have some mechanisms by which we can contribute to the national discussion.

[0:32:11]

Mark Elliot: Yeah, absolutely. I mean it's an interesting commentary on social science as much as the specifics of this particular area where the social sciences, as I think you said, is very good at sitting on the sidelines observing the world and making critical commentary about it. And what we're much less good at as a discipline, certainly in the UK, is getting creatively involved with stuff, sort of being involved at the beginning and then. So, it does, I think, actually bringing those skills to bear at an earlier point in process is a really important thing. And the work you were describing where you're engaging with public and thinking about how the tools should be developed is a really good example of that and we really need to see to see more of that I think if we're going to have what I call an active social science rather than a reactive social science.

So, any last thoughts before we close off then, Alex?

Alex Singleton: No, Mark, I think that's been a really great conversation. So again, just thank you for the invite.

Mark Elliot: You're very welcome.

[End of Transcript]