SOC ERC SYNDATA Candidate Information

Introduction
The PhD studentship will be focused on the social science study of synthetic data and will be based in the Department of Sociology at University of York, UK.

Studentship Provisions
The PhD studentship provides: Stipend at UKRI standard and full fee waiver; access to departmental training and module content (especially in qualitative research methods); the project's resources for other training opportunities, conferences, workshops, and summer schools; opportunities to join a supportive research culture within the department and for collaborating with others on the SYNDATA project.

About the SYNDATA Project
The SYNDATA project aims to better understand the ethical and political consequences of synthetic data for contemporary societies. Synthetic data are increasingly used to train and fine tune machine learning algorithms and generative AI models in a range of societal domains such as healthcare, finance, and government. Yet, they also promise to bypass a number of ethical issues related to machine learning and AI, such as gender and racial imbalances in training data as well as challenges of privacy and confidentiality in highly sensitive datasets.

The SYNDATA project will pioneer one of the first large-scale social science studies of synthetic data, and it will examine them in terms of three core areas or work packages: 1) ***extraction and generativity***; 2) ***diversity and difference***; and 3) ***bodies and resistance***. In order to do this, the research team within the project will conduct both archival research into the historical antecedents of synthetic data as well as qualitative research, such as digital ethnography and semi-structured interviews, into path-defining studies of different areas where synthetic data is currently being generated and deployed as a way to train algorithmic systems.

Your PhD Project within SYNDATA
The PhD student on the SYNDATA project will, together with the principal investigator (Benjamin Jacobsen), co-lead, co-design, and conduct research on the following work package: ***Diversity and difference***.

As part of this specific work package, the PhD and PI will analyse the ways in which synthetic data are generating societal differences as well as a new ethics of diversity. As algorithms have become an inextricable part of society, they have also been understood as technologies of exclusion and segregation that reinforce stereotypical and culturally entrenched representations in biased outputs based on characteristics such as race, class, and gender. Synthetic data constitute a shift in that they embody an explicit claim to actively generate diverse data points, such as images or text data representing racialised minority populations in, say, a healthcare dataset. The research within this work package will therefore focus on how the algorithmic generation of synthetic data is reconfiguring societal notions of difference and diversity. It will also explore the wider implications for social science notions such as population, bias, and discrimination. In this way, the project will contribute novel ideas and notions to debates around the ethics of algorithms and data, especially in terms of questions of difference, diversity, and representation.

In this work package, we will seek to respond to two main research questions:

Research Question 1: What is the relationship between the emergence of ‘diverse’ synthetic data and traditional statistical approaches to population?
There is already substantial social science literature that maps the histories of population-level statistics and their relation to issues such as governmentality, control, and power. There is also scholarship on the ways in which machine learning algorithms and AI are unsettling statistical ideas – such as normal distributions, regularities, and probability estimations – and generating new forms of racialisation and biased outputs based on the patterns learned in previously unseen input data. Synthetic data are intervening into such traditional statistical approaches. By representing diverse and rare examples (such as, for example, the faces of minority ethnic populations in a biometric dataset), synthetic data embody a claim to ‘balance out’ a data distribution and to ‘stretch out’ the tails of a population. This research will examine how synthetic data are generating new notions of (diverse) populations, drawing on ideas from the histories of statistics as well as critical data and algorithm studies.

Research Question 2: What is the impact of synthetic data and its claim to diversity on the wider social science vocabulary such as bias and representativeness?
In the computer science literature, there is a recurring slippage between the capacity to generate ‘diverse’ synthetic data points and the claim to resolve issues of algorithmic bias and representativeness. One of the reasons for this is that computer scientists have often framed the issue of, say, racial bias in facial recognition as an issue of creating more racially balanced (and hence, more ethical) training datasets. Synthetic data encapsulate a continuation of this logic: algorithms trained on diverse synthetic data points are often assumed to be less risky and less prone to prejudicial outputs. As such, the research will examine the ways in which synthetic data are intervening into and unsettling some of the core notions of AI ethics, such as algorithmic bias and representativeness.

The PhD and PI will do qualitative fieldwork research within AI research laboratories in the UK and in Europe. This qualitative fieldwork includes semi-structured interviews, non-participant observations, as well as ethnography. Interviews will be conducted with data scientists and machine learning engineers, and we will conduct ethnographic observations of the experimentations and deployments of algorithmic models where synthetic data is being used.

In addition to co-leading the work package, the PhD student will also collaborate with the rest of the SYNDATA research team on other relevant areas of synthetic data, participating in research meetings as well as in collaborative writing for publication.