Ada Lovelace Institute

Ethics and accountability in practice

Looking before we leap

Expanding ethical review processes for AI and data science research

Case studies

December 2022

Purpose of these case studies

Context

The Ada Lovelace Institute, the University of Exeter's Institute for Data Science and Artificial Intelligence, and the Alan Turing Institute developed these mock AI and data science research proposals for a workshop held as part of the research for their report *Looking before we leap*.[1] The workshop found that case studies are useful training resources for understanding common ethical challenges in artificial intelligence (AI) and data science research.

We are grateful to workshop participants for their contribution and feedback to the development of this case studies resource.

How should these case studies be used?

In our report, we found that academic and corporate Research Ethics Committees (RECs) are struggling to review the full set of ethical challenges that AI and data science research can pose. One concern that members of RECs highlighted was a lack of training materials that touch on the kinds of ethical challenges AI and data science research can pose. To that end, we have developed this set of 6 case studies that represent hypothetical submissions to a Research Ethics Committee.

These case studies are for use by students, researchers, members of research ethics committees, funders and other actors in the research ecosystem to further develop their ability to spot and evaluate common ethical issues in AI and data science research. Their purpose is to prompt reflection on common research ethics issues and the societal implications of different AI and data science research projects.

1 Ada Lovelace Institute. (2022). *Looking before we leap: expanding ethical review processes for AI and data science research*. Available at: https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

Source material

The case studies follow the template provided in *The Turing Way*,[2] with five broad questions on:

1. Project description
2. Data & methodology
3. Consent
4. Privacy and security
5. Further societal consequences

The case studies are entirely hypothetical; they do not reflect real projects, and any relation to an actual project is entirely incidental. This is to avoid subjecting any real project and its researchers to undue scrutiny. This comes at the cost of scientific feasibility. We – the authors of these hypothetical case studies – request that you look past the precise methodologies and the tone of the text, and focus on the potential substantial ethical issues these projects may raise.

General prompt

When reading through a case study, imagine that you are a member of a Research Ethics Committee. It is your job to ensure that the project meets the highest ethical standards. Keep the following questions in mind:

1 What potential harms does this project pose, both to participants but also to members of society who may be impacted by this work?

2 What measures should be put in place to mitigate against these risks?

3 What additional information do I need, and who should I speak with to find it?[3]

There is no 'right' answer to a case study - rather, this is an exercise to encourage reflection and discussion with your research group and peers.

2 The Turing Way Community et al. (2019). *The Turing Way: A Handbook for Reproducible Data Science*. Available at: https://the-turing-way.netlify.app/welcome

3 Specific prompts for each case study are available in Appendix 1.

A Google Form is available on the Ada Lovelace website[4] to allow you to review the case studies alongside the general prompt and receive a copy of your responses via email.

Questions? Contact Andrew Strait, Associate Director (Research Partnerships).

4 See: Ada Lovelace Institute. (2022). Looking before we leap: Case studies. Available at: https://www.adalovelaceinstitute.org/resource/research-ethics-case-studies/

Case study 1: Chemical misuse

Research Ethics Committee application form

1. Project goal and purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives.

This project aims to understand the internet-based spread of misinformation about the medical properties of an industrial grade chemical compound, sodium chlorite, which is being promoted or experimented with by different kinds of individuals and organised groups. Some promote regular intake as a form of 'homoeopathic' regime to prevent illness, but the compound has been promoted as a panacea to infectious diseases such as COVID-19 and malaria, as well as somatic cancers and many other diseases. In the vast majority of documented cases, intake strategies end up delivering harmful intoxicants to the body (of oneself, loved ones or strangers – including many vulnerable people). There have been many documented cases of poisoning, accidental death, harm warnings issued by the Food and Drug Administration, convictions from the US Department of Justice for sale of chemicals under false and harmful claims, and sale bans from European countries such as Ireland, to cite a few.

It is a global phenomenon including a particularly concerning spread in developing countries in Sub-Saharan Africa through organised local promotion. The harm that uncontrolled use of sodium chlorite homebrews is causing is alarming and we believe it to be on the rise.

The scientific community has pursued the development of drugs based on sodium chlorite (crucially, in the context of very specific therapeutic applications, delivery modes and extremely strict chemical production standards). Experimental results have been inconclusive at best, yet there is reason to believe that scientific activity on the compound could be adding to the hype. Myths of the medicinal properties of sodium chlorite have been circulating in the digital underworld for much longer. Practices of spontaneous consumption of industrial grade sodium chlorite include self-experimentation through quasi- or pseudo-scientific methods but also deceptive pseudo-religious rituals promoted by organised groups with financial conflicts of interest.

Practices of sodium chlorite consumption can be seen as a spectrum, from the respected to the discredited, from the scientific to the pseudo-religious, all of which is discussed and documented over the internet. The chemical is used in many different ways and as such is rarely if ever the 'same'. Discourses sometimes draw differences, and sometimes do not. Consumers usually, but not always, 'activate' sodium chlorite to transform it in chlorine dioxide before ingestion. The diversity of domains and styles of therapeutic experimentation is striking.

In the face of such a widespread interest and diversity of practices around the medical uses of sodium chlorite, we hypothesise the internet can make it easier for information fitting one's beliefs or interests to be strategically extracted from the social context of origination (e.g., scientific research), in order to be repurposed elsewhere, out of context, in a different domain.

In this project we want to understand:

1. How beliefs and evidence about sodium chlorite are communicated on the Internet.

2. The role (in discussions about homebrew consumption and the merits and properties of sodium chlorite) of scientific evidence (as well as scientific vocabulary, reasoning and theories) as well as opinions of authoritative leaders (scientists, politicians, etc.).

3. How peddlers, supporters and consumers of sodium chlorite justify their beliefs and try to convince others of them; what forms of evidence are corralled in conversations about sodium chlorite; what forms of evidential reasoning are employed.

We hope to be able to capture how much cross-communication there is between different domains, communities and websites; and specifically, what does travel well and what does not. For instance, how are sodium chlorite homebrewers following the unfolding of clinical experimentation? How do they come to misinterpret evidence of others' self-harm as evidence of therapeutic activity of the homebrew (as it has been observed they do)? What socio-cultural differences are relevant in these different online spaces (for instance, valuing anonymity, or dialogue, or argument; and norms for controversy management)?

The benefits this project aims to generate are:

1. Enabling better understanding of the way in which health misinformation spreads on internet platforms and forums, and the communicative and epistemic patterns that characterise it.

2. Better understanding of the styles of reasoning and standards of evidence employed by misinforming and misinformed individuals, and the travel of scientific evidence in the public domain.

3. Adding to the momentum, in the public sphere, needed to fight misinformation about sodium chlorite consumption and the harm that it can cause, by sensitising public opinion and informing policymakers about the complex communicative processes through which misinformation can confuse vulnerable individuals and put them in harm's way.

4. Informing regulatory responses as to how health misinformation aimed at industrial substance abuse should be tackled.

2. Data and research methods description

Please provide a description/overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

The project will be centred on computational social science work:

1) Data collection of textual communications (and visual attachments) spontaneously shared by individuals on web forums and platforms to discuss sodium chlorite and its health use; target forums and platforms will be determined after a qualitative scoping exercise but might include: Twitter, Reddit, MMSForum, NatMedHealth, Curezone and MammaMiaHealth.

Threads or selections thereof will be selected that discuss sodium chlorite use, its medicinal properties, visual evidence shared by the user and outward links to other webpages. Data will be collected through API where possible (Twitter, Reddit), or scraped when API is not available. Project partner MammaMiaHealth will directly contribute data from their own platform, and data scraped from third parties. The data we will use is publicly accessible and generated by users. Where possible, we will collect data dating from two years before, to two years after, a handful of key events that we have identified. Given the potential overlap, this might mean several consecutive years' worth of sodium chlorite communications data might be captured.

Key events include:

- the launch and completion of clinical trials testing sodium chlorite in patients of neurodegenerative disorders, between 2010–2012 and 2016–2017
- reporting of the Guardian newspaper on controversial pseudo-experimental activity by sodium chlorite peddlers in Sub-Saharan Africa (2019)
- the infamous White House press conference in April 2020 where Donald Trump suggested researchers should test if COVID-19 can be cured by injecting bleach.

The volume of these communications is expected to be very small relative to the size of some platforms. A preliminary scoping query on Reddit returned a dataset equivalent to 40k tweets mentioning Trump and bleach, of which only a small minority (if at all) could be discussing the event in relation to sodium chlorite, given the substantial mainstream media coverage of the event.

2) Data analysis including:

1. Network analysis: social network analysis, identification of strong and weak ties, influencers, etc. within and between websites – to understand link patterns between websites, and between people interacting on sodium chlorite.

2. Neuro-linguistic programming (NLP)-based vocabulary and argument analysis: analysis of linguistic and reasoning styles through 'argument extractors', of vocabularies of evidence, and reasoning eclecticism (drawing information from a heterogeneity of sources) – to understand how people talk about sodium chlorite and what conceptual and evidential resources they mobilise to speak their beliefs about it.

3. Event analysis: observing change in communications and patterns of social interaction before and after an event of interest, to understand how key events (publication of trial results, opinions of public officials) impact discussions and beliefs about sodium chlorite in online communities.

4. Misinformation profiling: NLP-based identification of the most active 'misinformants', i.e. people who engage in blatant misinformation

about sodium chlorite; analysis of their communication and interaction patterns, including evidential and argumentative styles, and fact-checking analysis of an individual's communications (project partner, fact-checking organisation Facts4all will collaborate on the definition of a consolidated fact-base). In addition to these NLP-based analyses, and in order to understand who are the individuals that 'bridge' between different communities and social spaces, we will analyse the digital presence of the same individual across different websites through probabilistic data linkage profiling (anonymised at the data collection level, further details below) and/or unsupervised clustering methods, such as latent Markov modelling. Project partner MammaMiaHealth will consult on this last analysis task, contributing methods developed during a first analysis they had preliminarily conducted.

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

We aim to collect data from publicly accessible websites or sources. An argument exists that views public open websites (non-walled gardens) as a form of public space, exempting them from explicit consent. Users are aware that their posts and their profiles are publicly available and easy to analyse by other website users. Communications that users wish not to be public can be easily deleted or transferred to private channels; private communication is always available to them. We are also aware that UK copyright law contains a provision that website data can be scraped

when it is only to be used for academic purposes – as is the case here.

Internet research ethical guidelines suggest that if explicit consent from individuals is not required then researchers must make sure that the risk of harm is minimised, the research delivers prosocial benefits, and users' anonymity and privacy is protected at all times. We aim to follow these guidelines and address these points in detail below.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

The data that we collect is not anonymised at the point of collection and protecting the anonymity of users is our greatest ethical concern. We are very aware of the risks in deanonymization given the sensitive subject matter; if we were to make our data available without any anonymization then users could be targeted by opposing individuals. To account for this, we will not present individual content (such as posts or usernames or metadata) in our results. All analyses and visualisations will be in aggregate, which will stop any individuals from being identified. When typifying styles of reasoning and communication, we will use paraphrasis.

No individually identifiable information will be made available to researchers outside of the research team for the duration of the project. Information will be constrained even for researchers inside the research team. The data will be anonymised upon collection through a split-file approach separating and encrypting identifying metadata (this method is inspired by more sophisticated and secure approaches, pioneered by

large public health infrastructures such as the SAIL databank) that is then stored in a separate encrypted drive. This will help to ensure that the communication of a user in the public space of a website remains anonymous to researchers – thus creating a similar condition to how strangers in a public space can access others' communications without, in most cases, knowing who each other is exactly. Researchers will then work on complex individual data where identifiable information is substituted by encrypted identifiers.

We will store processed data on fully encrypted disks and/or Spideroak's 'zero-knowledge' state of the art encrypted collaboration software CrossClave. Encrypted cloud storage will also be used to store the data whilst it is being processed - it won't be stored there for the long term. We will use the Azure cloud servers, which also use 256-bit AES encryption.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

Ethics/harm to researchers: We are aware of the potential of harm risked by researchers through exposure to disturbing and/or graphic content of self-harm and abuse. We believe that the overall risk to researchers is low because a large portion of the work is computational. Ensuring the safety and wellbeing of our researchers is a key concern and we will discuss this issue as part of our regular team meetings.

Minimising harm: The goal of this research is to improve our understanding of the social dynamics of online behaviour. At no point will we intervene in individuals' online environments or make direct contact with them. Our work will not lead to the creation of new classifiers or tools that could be repurposed for nefarious ends, such as social surveillance or repression.

Benefits: The potential harm caused by sodium chlorite misinformation groups is substantial and tangible, and has been linked to devastating offline incidents – including several deaths and injury. Most determined perpetrators have been on the run and international arrests have been carried out against them, underscoring the nature of public threat from the practices that we are trying to study. While we will do everything to minimise new and unexpected harms, it is crucial that we conduct research into this area to better understand the media ecosystem, their discourses and any behavioural changes.

Limits: We are aware that this research may not be generalisable across all sodium chlorite misinformation online and are presenting our results as a highly contextualised study of online behaviour and communication.

Case study 2: Synthetic iodide data

Research Ethics Committee application form

1. Project goal and purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives.

This project will be conducted as a hackathon. Its main aim is to provide participants (early career researchers) the opportunity to engage with real-life scientific research. The specific project question relates to the creation of synthetic data models for the representation of sea surface iodide concentrations.

The marine iodine cycle has impacted air quality and atmospheric chemistry. Specifically, iodide reacts with ozone in the top few micrometres of the surface ocean. This reaction leads to a substantive drop in tropospheric ozone (a pollutant gas) and is an important source of reactive iodine in the atmosphere. Sea surface iodide parameterisations are now being implemented in air quality models, but these are currently a major source of uncertainty.

During this one-week hackathon, we seek to develop an efficient, fast and scalable synthetic graph generation method which captures the relational structure and attributes of the original data collated by

Chance et al,[5] as well as data from a private corporation based in the USA, which will cover data collected since 2018. This will allow for the generation of more accurate parametric data. This data's increased precision will entail more accurate models for the prediction of changes in air quality, as well as future tropospheric ozone losses, 15% of which are accounted for by ocean surface-level iodide.[6]

This project aligns with the university's values by (1) providing learning opportunities to future researchers and (2) working on understanding the environment, a key aspect of the university's strategic plan: *2030 & Beyond.*

2. Data and research methods description

Please provide a description/overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

5 Chance, R.J. et al. (2019). 'Global sea-surface iodide observations, 1967–2018'. *Nature*. Available at: https://www.nature.com/articles/s41597-019-0288-y

6 Sherwen, T. (2016). 'Iodine's impact on tropospheric oxidants: a global model study in GEOS-Chem'. *Atmospheric Chemistry and Physics*. Available at: https://acp.copernicus.org/articles/16/1161/2016/

The data employed by this project has two sources:

1. Data on ocean surface iodide levels (iodide observations) from 1967 to 2018, made available by Chance et al,[7] and we will also use the metadata they have made available.[8] We will be using the data based on what the authors have said when making it available: 'the data may be used to model sea surface iodide concentrations or as a reference for future observations.'

2. Iodide observations since 2018 are provided by RCC LLC., a US-based firm that compiles large quantities of data to support merchant vessels. This data will be made available to us following the terms of a data sharing agreement, which is currently being negotiated by the Legal team. We will update the research ethics committee when this is finalised.

A first exercise for the hackathon participants will be to generate parametric data that can later train test models, although the focus is the accuracy of the parametric data. For the generation of the required training dataset, we will employ a generative adversarial network (GAN). In the GAN model, two networks, generator and discriminator, will train a data-creation model iteratively. The generator takes random sample data and generates a synthetic dataset. The discriminator compares synthetically generated data with the real dataset. Python is used for setting up the simulations and saving the results.

The generated synthetic data is then used in commonly used models for the prediction of changes in air quality and future tropospheric ozone losses. We will compare the results of these models when using the original dataset versus our generated data.

7 Chance, R.J. et al. (2019). 'Global sea-surface iodide observations, 1967–2018'. *Nature*. Available at: https://www.nature.com/articles/s41597-019-0288-y

8 Chance, R.J. et al. (2019). 'Metadata record for: Global sea-surface iodide observations, 1967-2018'. *Nature*. Available at: https://doi.org/10.6084/m9.figshare.10130129

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

No data employed by this project pertains to individual persons. In reference to the use of existing data, see above.

The data employed by this project is made available by Chance et al and RCC LLC.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

The data we use does not raise security concerns. We will employ the university's IT infrastructure, as provided for the purpose of these hackathons. The university also has protocols in place for the maintenance of data beyond the completions of such hackathons. These will comply with RCC's data sharing agreement.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

This project raises four ethical concerns relating to (1) the use of synthetic data, (2) science communication, (3) environmental impact and (4) intellectual property.

On the use of synthetic data, the principal investigator and more senior members of the research team will inform participants of the limitations of synthetic data. Indeed, there is an inherent 'probabilistic looseness' when fitting synthetic data to models about the real world. The team will work on a joint statement to emphasise this point when introducing the project to the hackathon participants. Acknowledging the limits of synthetic data is what prompted the need to test the data on numerous models at the end of the project.

The point about science communication is deeply interlinked. Indeed, the limitations of the data mean that findings – the parameters for synthetic data – must be cautiously communicated as tested solely on models. The project will not claim to have developed data that can be employed for the analysis of ozone-iodide interactions for the prediction of air quality and ozone losses. The final findings will be communicated solely as pertaining to the models employed for testing the robustness of the generated synthetic dataset.

In either case, it is worth noting the benefits of such a dataset. The synthetic data responds to the need to better understand environmental challenges so relevant today, before the advent of global warming. This very important context also helps us show caution in our claims. Ultimately, developing synthetic data that could be incorrectly used in prediction models they have not been tested on can lead to false findings. With this risk in mind, the synthetic data we develop will come with metadata and documentation showing the methodology for its creation, emphasising the limited models it has been found to work in.

On environmental impact, it cannot be ignored that large computational models involve high-energy consumption[9] that can itself be detrimental to the wellbeing of the environment.[10] The use of AI for the study of the environment seems counterproductive. And there is indeed a risk for data such as what this project produces to be used in resource-intensive models. We feel it will be out of our hands to ensure that others use our resulting dataset on Green500[11] computers, for example, but we will provide a statement encouraging the thoughtful use of our data as documentation when sharing the results publicly.

Regarding intellectual property (IP), the hackathon framework at the university has special provisions for how we approach data providers – in this case, RCC LLC. The data sharing agreement will most likely result in either (1) our being able to share their data and our findings or (2), our being able to share only findings without reference to their data. RCC LLC are aware of the public good we seek to provide through these

9 Bender, E. M. et al. (2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, (FAccT '21)*, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922

10 Coeckelbergh, M. (2021). 'AI for climate: freedom, justice, and other ethical and political challenges'. *AI and ethics*, 1, pp. 67–72. Available at: https://link.springer.com/article/10.1007/s43681-020-00007-2

11 The Green500 is a biannual ranking of supercomputers, from the TOP500 list of supercomputers, in terms of energy efficiency.

hackathons, but it is likely that the IP of their datasets means that these cannot be shared – they are resource-intensive to gather, and fulfil a goal within RCC's business model. It is therefore likely that we end up in scenario 2. A question then arises about the reproducibility of the project's outputs. In scenario 2, it is unlikely that sufficient data will be made publicly available for other researchers to inquire deeply into our project. To this effect, we return to the special provisions for hackathon data providers such as RCC: the data sharing agreement should allow for data to be made available to researchers who wish to reproduce our findings. However, at this point, these are just conjectures and we will update the Research Ethics Committee when the data sharing agreement is completed. If the negotiation falls through – a possibility – we will unfortunately have to go ahead without RCC.

Case study 3: Spiking Neural Networks

Research Ethics Committee application form

1. Project goal and purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives.

Background: Neural networks have become the key technology of AI in the form of Artificial Neural Networks. Research interest in neural networks has grown in the last few years due to their success in terms of accuracy in applications such image classification, multiple object detection, language translation and speech recognition. More recently, interest in the research community has increased in Spiking Neural Networks (SNNs) due to their rich temporal dynamics and high-power efficiency. SNNs emulate the structure of the nervous system as a network of synapses that transmit information via ion channels in the form of action potential, or spikes, as they occur. SNNs have shown great potential to achieve advanced robotic intelligence in terms of speed, efficiency and computation abilities.

Purpose: The purpose of this project is to understand the performance of SNNs for the application in neurorobotics by evaluating their performance in triggering finger motion reflexes on a robotic hand. The main motivation of this work is to control a robotic hand with human muscle signals using a Spiking Neural Network (SNN).

Intended benefits: The intended benefits are to enhance understanding of SNNs and potential applications for more human-like robotics that enable more natural and responsive human-machine interactions. More specifically, the interaction between robots and humans is of great importance for the field of neurorobotics as it can provide insights on how humans perform motor control and sensor processing, and on how it can be applied to robotics.

2. Data and research methods description

Please provide a description/overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

To test the performance of Spiking Neural Networks (SNNs), the first part consists of collecting human surface electromyography (sEMG) to measure human muscle activity and encoding these signals into spikes. The second part involves the SNN performing classification of the data to detect which fingers are active in the human hand and generating an activation signal. The activation signal is used to trigger a motion reflex using a motor primitive, which is then mapped onto the robot kinematics. Finally, the SNN is evaluated by participants wearing a sEMG sensor and recording a training dataset, and flexing different fingers.

The novelty of this approach is that the classification and generation of a motor primitive is implemented using SNNs. The notion of motor primitives relies on the assumption that the human motor system

functions by the central nervous system using different base motor components in a hierarchy. Therefore, the concept of SNNs simulates the classification and generation process of movement in the human brain. This could therefore be regarded as an initial model of this biological process.

The SNNs will be evaluated by users wearing a non-invasive sEMG sensor, recording a training dataset, and flexing different fingers. The muscle activity is recorded using a Myo sensor. The sEMG signals will be encoded into spikes as input for the SNN. It will be tested whether the classification can detect the active finger and trigger the motion generation of finger reflexes. It will be evaluated whether the SNN is able to control a real Schunk SVH 5-finger robotic hand.

Overall, a sample of 200 participants will be recruited, which will wear the gesture control Myo sensor to record sEMG data. The participants will be recruited among physically healthy volunteers. The Myo sensor is a commercially available, proprietary sensor produced by Thalmic Labs, a small start-up based in the UK, and commonly used by hobbyists in the context of robotics, drones or games. The company provides open APIs and free SDKs and is therefore easily accessible by researchers. It is made up of eight blocks with non-invasive sEMG sensors that provide data on electrical activity. The armband is used around the middle of the forearm. When a finger is moved, the muscle's electrical activity is recorded using the eight different sensors. The sensor has an indicator so that it can always be placed in a similar way. In order to record consistent data with the sensor, the segment with the LED light has to be placed approximately at the same position. We use a marker to mark the position on the underarm skin.

For each user a training dataset is required with multiple samples. A sample consists of a continuous sequence of finger activation in one hand and each finger has to be flexed for a short period of time. This procedure is repeated several times. The training data has to be recorded as a continuous sEMG stream of all eight channels with appropriate binary labels for the time during which a finger is pressed.

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

The participants will be informed about the purpose and objectives of this research prior to the test. Participants will receive an information sheet, detailing the precise purpose and benefits of the research, the tasks they have to perform, any potential risks involved, and be informed that their participation is voluntary and that they can withdraw from the research at any time without giving a reason. They will further be informed about how their data will be collected, stored and used for the purpose of this research. Furthermore, the participants will have to sign a consent form before commencing the study.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

Although every effort is made to secure the data, as the sensors are proprietary, the data will be stored on the manufacturer's server – to which the research team has password-protected access. The data is stored using the manufacturer's internal security standards and based on their own data storage policy. We are working with the manufacturer to provide shareable versions of the data that does not rely on the data being stored on their server. However, this is rather problematic as the company is a small start-up that does not routinely share data this way, and does not have the infrastructure to do so. The dataset, however, does not contain any personally identifiable data, and should not form any obstacle in terms of privacy or consent.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

The research team does not anticipate that any direct harm to the individual participants would arise from the research. As the data collected pertains to muscle movement, there is no risk around issues of privacy, and participants give consent for their data to be used for the purpose of this study only. There is a small risk that the data may lack accuracy, but through taking care to always place the sensor in the same area of the arm, and repeated tests and multiple samples, this risk is minimised.

The research relies on the notion of motor primitives, and the concept of Spiking Neural Networks (SNNs) simulates the classification and generation process of movement in the human brain. While SNNs are simulating brain function and the nervous system, these are in the early stages of research and the concept of SNNs may be prone to error.

Research into powerful SNNs and, in particular, large datasets of human muscle movement, may be accessed by third parties, and used to control other technology, such as drones.

Case study 4: Tattoo ID to map gangs

Research Ethics Committee application form

1. Project Goal and Purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives

Mapping London's gangs through a deep-learning tattoo image classification model

Purpose and Goal

The purpose of this project is to train an image-based classifier that can categorise different tattoos that are associated with various London-area gangs. When shown an image of a tattoo, the model will be able to predict what, if any, gang the owner of the tattoo is associated with, helping local police departments better assess a suspect's potential connection to organised crime and enabling safer prison system allocation by preventing members of rival gangs from being unknowingly housed together.

Throughout the last 10 years, London has seen a surge in gang activity. According to the Metropolitan Police Service (the Met), knife crime is at

its highest point in the last 10 years,[12] graffiti and public vandalism crimes have increased by 30% since 2016 and petty thefts are up by 42% since 2015. The Met blames this increase of criminal activity on a new wave of organised criminal organisations, many of which are engaged in far more serious crimes relating to drugs, trafficking and murder. This increase in gang activity has come at a time of decreased costs for the police service.

According to police sources interviewed for this project, one of the greatest needs police departments currently face is tracking who is involved in which gang, which may help investigators develop a map of collective criminal activity and associations. Similarly, Her Majesty's Prison and Probation Service (HMPPS) reports that a critical failure in ensuring safe outcomes for their inmates is not knowing what gangs their inmates are affiliated with. This creates a severe risk of inmates from rival gangs being housed in the same cell block or cell, imposing a greater risk on their lives and of inmates around them.

According to our police informants for this project, tattoos are one of the most reliable sources of information about an individual's gang relationships. Tattoos can reflect a gang's location, name and even specific information about the role the individual plays in the gang (e.g. as an enforcer, drug dealer etc.). HMPPS and the Met already maintain a large database of high-quality tattoo image data collected during arrests, surveillance operations and incoming inmate interviews. By training an image classification model on this data, we can help police and prison officers perform their job more effectively and at a higher quality.

Intended Benefits

Our goal for this project is to create an image classification model which uses non-personal data and helps make the job of police officers and prison officers easier and more effective, while also ensuring the safety of inmates.

12 Allen, G. and Harding, M. (2021). Research Briefing - Knife crime statistics. House of Commons Library. Available at: https://commonslibrary.parliament.uk/research-briefings/sn04304/

2. Data and research methods description

Please provide a description/overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

This project will involve training an image classification system on a database of tattoo images that are known to be affiliated with particular London gangs. This database, TATDAT, began as a joint effort between the Metropolitan Police Service (the Met) and HM Prison and Probation Service (HMPPS) in 2009. The database contains over 1.3 million photos of tattoos taken by officers and prison guards during arrests, surveillance operations and pre-incarceration inmate interviews.

The images in this database are mostly high resolution close-up images taken of a specific tattoo, with no other features of the individual available. Some images are taken during surveillance operations from a distance with a high-powered lens and, while grainer in nature, they also focus on the tattoo. No images contain identifying features of the person themselves, apart from a small subset of tattoos that are of the face (we have excluded these from our dataset). The images are not linked to a name or any other personal identifying information, but are linked by a unique personal identifying code (UPIC). UPIC 4, for example, would include pictures throughout the database of their tattoos, including any changes over time the tattoos may have undergone. The only other information attached to the UPIC is a 'gang affiliation' category, shorthand: GANG_AFF. This information is input by the arresting officer or the prison officer responsible for taking the relevant photo. There are over 83 different gangs listed in the database. Tattoos that are not confirmed to be associated with a gang are listed as 'none' in GANG_AFF.

This data was created for the purpose of capturing known gang tattoos. It is currently used by police and prison officers to manually comb through its database and match any newly suspected gang tattoos to known tattoos in the database. This is an extremely time intensive process – one officer stated this can take upwards of 14 hours of officer time per suspect; time which could be spent investigating cases.

Our method will involve using a machine learning image classification model trained on a subset of these categorised images. The classification model will be able to take new images uploaded to the TATDAT database and predict the likelihood over the new person's gang affiliation status.

Future research might also seek to predict the specific role that a gang member plays based on the nature of the tattoo. As noted earlier, some tattoos signify particular activities – for example, a small bullet underneath the eye may signify that the owner of the tattoo is a hitman. We acknowledge that our model will not be able to make these classifications at this time, but we understand the owners of the TATDAT database plan to include information on the suspect's charge, starting in 2022, which may be useful for predicting their role in a gang.

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

We appreciate this project may raise some concerns around consent. The database that we are using is lawfully collected by the Metropolitan Police Service (the Met) and HM Prison and Probation Service (HMPPS) with the intention of being used to aid law enforcement and prison officers with identifying the potential gang affiliation of a suspect. All images were collected during lawful police arrests, surveillance operations and prison inductions. The data does not contain personal identifying information insofar as the images are solely of the individual's tattoos with no additional information. In our eyes, we are only helping police and prisons officers conduct a lawful activity more effectively and efficiently.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

The TATDAT database is only accessible by approved Metropolitan Police Service (the Met) and HM Prison and Probation Service (HMPPS) officers. It is stored on an encrypted cloud service run via the National Crime Agency. This project has been sponsored by the Met, who have agreed to allow our research team access to this database on the condition that we perform all computational training on-site at their research department headquarters, on approved police service laptops.

Since the data is only of an individual's tattoo and gang affiliation with no other demographic information or personal data, and since we have removed any images that include identifying features such as their face from our dataset, we do not see any concerns around the need to further anonymise an individual.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

We appreciate that there is a growing debate around the role of policing in contemporary British society, with real and proven concerns of racial bias and profiling in policing. While we share these concerns as researchers, we also appreciate the need for law enforcement to do their job effectively given the increasing rates of gang activity combined with significant cuts in the policing and prisons budget. It is our intention to help keep people safe but we must grapple with the reality of biased data that is rife in the criminal justice and law enforcement sector.

A weakness in our data is that determinations of whether a person is actually affiliated with a gang are entirely dependent on the judgement of the underlying officer who tagged this data, a judgement which we are unable to contest or edify. This raises a serious epistemic concern

that our prediction of someone's likelihood of being in a gang will be interpreted as 100% fact, when it is merely a prediction based on potentially historical data. In cases of a false positive, we see some risk of harm. A person may be affiliated with a gang, which may affect the severity of their sentencing or lead to increased police time investigating their connections. In cases of a false negative, we see some risk to a potential inmate being unknowingly transferred to a penitentiary with rival gang members present.

To address this challenge, we will include an explicit warning label in every prediction we make that this model is not 100% accurate. We will follow up our testing with live trials of this system in use, which will help give us a ground truth of whether this model's prediction aligns with the actual affiliation of new inmates and arrestees. In testing our model, we will seek out grey-area tattoos (e.g. those of barbed wire, or geometric signs) that may in some cases be affiliated with a gang affiliation while in other cases associated with a popular innocuous tattoo design.

Overall, we believe our tattoo classification scheme will do more good than harm. This system merely speeds up a process that humans already do, and still retains a human decision-maker at its core.

Case study 5: Emotion recognition

Research Ethics Committee application form

1. Project goal and purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives.

Purpose and goal: The purpose of this study is to develop a software in the form of a game with the aim to aid the development of recognising emotions in social interactions for autistic people.

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised by restricted and repetitive behavioural patterns, impairment of social interaction and verbal and non-verbal communication such as facial expressions, gestures, and eye contact, as well as socio-emotional reciprocity deficits present since the beginning of childhood. The ability to recognise emotions is particularly compromised in individuals on the autism spectrum.

There are already a number of assistive technologies and software programmes that have been evaluated in terms of emotion recognition in the context of ASD. However, existing tools rely on emotion recognition through, for example, training the recognition of photographed or filmed faces expressing emotions, such as the prominent Cambridge Mindreading (CAM) Face-Voice Battery test. None of the available tools

make use of non-invasive, automated emotion recognition using facial recognition technologies. This is therefore a pilot study, in which such a software will be developed and tested for its accuracy and suitability as a learning tool for autistic people.

Benefit: The benefit of an emotion recognition tool is that it could help autistic people to recognise their own emotions, while also learning to understand expressions in others. This would significantly help autistic people to increase skills in social interactions and communication.

2. Data and research methods description

Please provide a description/overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

As part of the study, a software tool will be developed and evaluated that uses automated facial recognition, based on algorithms trained on existing emotion recognition datasets. A common emotion recognition model is the categorical model which uses six basic emotion categories (happiness, anger, surprise, sadness, disgust and fear). This is based on the notion of these emotion categories being universally recognisable and 'hard-wired' through evolutionary processes into the brain. Another perspective on emotions is the dimensional model, which demonstrates that emotions can occur along the dimensions of arousal and valence, also termed the model of 'core affect'. Both models have been used to train and classify emotions, and to create emotion recognition datasets.

While earlier datasets relied on posed behaviour in controlled environments, more recent approaches recognise that emotions occur in natural and uncontrolled contexts, such as the annual 'emotion recognition in the wild challenge' as part of the ACM conference on multimodal interaction. In this context emerged the Static Facial Expressions in the Wild (SFEW) dataset, which contains 700 face images, based on 37 movies, which have been manually annotated using the six basic emotion categories as discussed earlier. Other databases rely on the dimensional representation of emotions, such as the SEMAINE dataset. For example, the Aff-Wild dataset uses emotion representations based on 298 videos taken from YouTube that have been manually annotated using the categories arousal and valence. The AffectNet is the largest database based on querying different search engines using 1,250 emotion-related keywords in six different languages. The database contains over 1 million images of emotional facial expressions, which were manually annotated using both the categorical and dimensional models of emotions.

The software will be developed and the algorithm will be trained on a combination of these most recent databases (SFEW, SEMAINE, Aff-Wild and AffectNet) as these are most suitable for determining emotions using facial recognition technology. The aim is to develop an emotion recognition system that is highly accurate and based on objective measurements of the users' affective states.

The software will emulate a social situation in the form of a game, which will enable the user to learn about the role of emotions in social interactions.

The software will recognise emotions based on facial recognition technology and the user learns about their own emotions through feedback in the game. Other non-player characters in the game are designed to respond to the user and adjust their behaviour based on the emotions that are recognised in the user. For example, if the user knows that they are 'friendly', other non-player characters will approach; when the user is 'angry', the characters will withdraw. This is designed so that the user understands their own emotions and why other actors in the game respond in a certain way.

The emotion recognition software is tested on participants with a diagnosis of autism. For this study, 8 participants will be recruited and invited to interact with the game in three two-hour sessions.

The effectiveness of the game will be evaluated using a Tobii eye tracker to measure attention and engagement levels during gameplay, and a questionnaire is provided after the game is completed. The questionnaire evaluates whether the participants developed an increased awareness of their own emotions, and how these may influence other people's behaviours in social interaction.

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

The participants will receive an information sheet with details on the purpose of the study, benefits of participating, potential risks and be informed about consent. They will be advised that their participation is voluntary and that consent can be revoked at any point without giving a reason. The participants will also be informed that their data is confidential and will be anonymised, where the data will be stored and how it is going to be used.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

For the Tobii eye tracker, the data is recorded and processed through the company's integrated Tobii Pro Lab, which is a software specifically designed to automate the process of data analysis and presentation of results. The researchers have access to the raw data and the dashboard at any time through a password-protected gateway. The data from the questionnaire will be stored on a secure and encrypted cloud-server owned by the university, to which only members of the research team have access through their university ID.

The data from the Tobii eye tracker, and the questionnaire are anonymised through pseudonymisation, where any personally identifiable information will be replaced with a code. The identifying data will be kept on a university-owned, encrypted hard drive to which only the research lead has access. This file will be discarded at the end of the research project. The anonymised research data cannot be used for the purpose of identifying individual participants by triangulating with other publicly available datasets.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

Potential risks of the study could include that the datasets that train the algorithm of the emotion recognition system could be inaccurate, especially since the datasets are manually annotated using basic emotions and the dimensions of arousal and valence in images or videos sourced from the internet.

Another risk is that the emotion recognition software may distort the participants' perception of emotions, in particular if the system is inaccurate. This may result in further confusion in autistic people and affect trust in their own ability to recognise internal affective states.

Finally, there is a risk that these systems may be deployed in other contexts, such as emotion recognition in clinical contexts, e.g., monitoring compliance with treatment or accessing clinical care based on AI-powered systems.

Case study 6: VPN app

Research Ethics Committee application form

1. Project goal and purpose

Please introduce your project.

Tell us, among other things:

- the purpose and goal of your project
- the intended benefits of your project
- how this research project aligns with the university's goals, challenges, and/or objectives.

Purpose and goal: Social media platforms have come under scrutiny for their misuse of personal data (e.g., the Facebook-Cambridge Analytica scandal),[13] for nudging individuals at the expense of their freedom[14] and even censoring activists.[15] Meanwhile, the algorithms that govern such sites are proprietary – they are owned and protected by large corporations. Furthermore, even if they were open source, the precise manner in which they 'make decisions' will be like a 'black box' – too opaque to be interpretable.

This project will have research participants install a particular virtual private network (VPN) service on all the electronic devices they use to access social media. Whilst this VPN service will provide the basic protections we expect from a VPN, it will be enhanced with particular

13 See: BBC News topic 'Facebook-Cambridge Analytica scandal'. Available at: https://www.bbc.co.uk/news/topics/c81zyn0888lt

14 See: Mitchell, L. and Bagrow, J. (2020). 'Do social media algorithms erode our ability to make decisions freely? The jury is out'. *The Conversation*. Available at: https://theconversation.com/do-social-media-algorithms-erode-our-ability-to-make-decisions-freely-the-jury-is-out-140729

15 See: Lim, M. and Alrasheed, G. (2021). 'Beyond a technical bug: Biased algorithms and moderation are censoring activists on social media'. *The Conversation*. Available at: https://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669

usage logs. These logs will secure data about the posts and users who are amplified on different platforms, as well as what content is censored. These logs will constitute the main source of data for the research team to analyse.

Benefits: The benefit of gaining insight into algorithmic decision-making processes on social media platforms is that researchers, citizens and activists can be informed about the precise societal impacts of online algorithms. It can also help shape public policy concerning social media and broader internet regulation.

Institutional Alignment: This project sits within the institution's efforts in providing algorithmic audit solutions.

2. Data and research methods description

Please provide a description / overview of the data that you will use for your project, if any, as well as the research methods you will use.

Please cover:

- the type of data
- the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data
- the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from
- where and how the data to be used was collected (if known)
- a brief description of the research to be carried out and the methods that will be used.

The project will follow three main stages: (1) the design of the VPN app, (2) the recruitment of research participants to install the app and (3) the analysis of data collected by the apps. At different stages, different expertise will be required, as described below.

Stage 1: the VPN app

This stage has involved the principal investigator, two software engineers and one data analyst. The app has already had its major features designed and is currently at its development stage. We expect it to be fully deployable within one month.

The end goal of the app was to develop something that could monitor how social media algorithms respond to different users and their engagement with those platforms. The app could have been designed in at least two ways. For example, it could have been a web browser (like The Markup's).[16] However, this would require research participants to access social media through the new browser rather than the social media apps that would normally be used. This could risk affecting their way of navigating the platforms, but also missing out on useful data if the participants employed the usual social media apps if they, say, forgot about our browser.

The second way to develop the app was to capture as much data as possible about social media algorithms. This meant developing something that could collect data from various social media apps. We found the 'VPN' option to be the most effective, as it could collect data from all interactions the device has online. For example, the app can inform the researchers – who see data streaming live – how a user performs across different platforms and according to their different proprietary algorithms.

Stage 2: research participants

This stage will involve the principal investigator and a research assistant with a social sciences background.

This project will follow the institution's guidelines on research participant recruitment, as well as employing the standard personal information sheet and consent form provided.

16 See: The Markup. (2020). The Citizen Browser Project—Auditing the Algorithms of Disinformation. Available at: https://themarkup.org/citizen-browser

The personal information sheet will inform the participants of the following particulars, also pertinent to our methodology:

1. Participants will be required to download our VPN app and have it activated for the duration of the study, which is three weeks.

2. The app will provide data pertaining to the participants' interactions with specific social media platforms, including Facebook, Instagram, Reddit, TikTok and Twitter.

3. This data is sent live to our secure IT infrastructure, screened by an algorithm to ensure the data we store is relevant (for example, activity on certain apps e.g., Facebook Messenger or WhatsApp, will be deleted immediately).

4. No personal data will be collected throughout this period as the app only gathers data about a device and its interactions with online sites, with zero risk of accessing the camera, photos or microphone.

5. After the three weeks, the app will cease to work and participants will be prompted to uninstall it. For security, participants should reinstall the VPN they were using before this study.

Stage 3: data analysis

This stage will involve the principal investigator, the research assistant and two data analysts.

Data will be compiled during the three weeks when research participants have the apps installed. Data analysis can begin during this period, especially to ensure that the VPN apps are functioning and the participants using them appropriately.

An initial process of data cleansing will be conducted, as well as data labelling. During the first two weeks after Stage 2, any data that should not have made it to our servers will be deleted. For example, activity from a site that the screening algorithm inaccurately classified as a social media platform will be immediately deleted.

We are still unclear as to what precise methodologies for data analysis will be employed at this stage (we hope the Research Ethics Committee has some thoughts on this).

3. Consent

Please comment on any issues around securing consent raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If your study involves collecting data about, with, or from human participants, will they receive all relevant information about what their participation will involve, prior to providing consent?
- If your study involves using existing data about people, was their consent given for the original data collection? Did they consent to this data being reused?

Consent will be captured from research participants following our standard protocols. Participants will need to have provided consent the week prior to Stage 2 to continue being a part of the project and be given access to the VPN app.

Although no personally identifiable data will be gathered, we believe the simple involvement of the participants during the three weeks is enough to call on our duty of care towards them.

Furthermore, the data will originate from their personal devices, and we – the research team – believe that this renders the data of a personal nature.

4. Privacy and security

Please comment on any issues of privacy and security raised by your research.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- If you use data about people in your study, is this data anonymous, or anonymised? Could it be associated with identifiable individuals, including by triangulating with other publicly available datasets?
- How do you plan to keep sensitive data safe and secure? What will you do with sensitive data after the study is completed?

Data from the VPN app will be fed directly into a secure IT environment hosted by the institution and meeting our standard security protocols.

It is important to note that, whilst data will not be personally identifiable, it can be used to re-identify individuals by triangulating with public social media activity. This is the main reason why the data will be stored securely and accessible only to the pertinent research team members at each of its stages.

When data reaches the IT environment, furthermore, it is automatically screened by an algorithm, as mentioned when describing Stage 2 in section 2 above. This is to mitigate against the risk of collecting too much data and to uphold the GDPR's principle of data minimisation. A further protocol is in place for this process, with data being deleted under the same circumstances by a human analyst at Stage 3 (also mentioned in section 2 above).

The main risk to privacy and security is if a researcher with access to the IT environment was to either leak data or store unnecessary data, resulting in analyses that involved sensitive or private information from apps that are not social media platforms. We believe the institution's whistleblowing and disciplinary policies are sufficient to hold individuals accountable, and that the research ethics training provided for all researchers encourages integrity.

5. Other harms

Please comment on the potential for individual, societal or ecological harms to arise from your research, beyond what is described above.

Where you identify risks, you should inform us how your research plans to minimise or eliminate them.

Questions to consider:

- Could any harms arise to the people involved in conducting this research (e.g. viewing violent content could harm researchers)?
- Could conducting or promoting this research create unintended negative outcomes, such as environmental damage, new power imbalances or the misuse of technology?
- How do you plan to ascertain and acknowledge the limitations of your research, if any (e.g. does the data sample you use allow for generisability of your research findings)?
- What benefits could your research contribute that would balance or outweigh any potentially negative impacts that could arise?

We are aware of three areas of concern and have implemented measures and conducted analyses to mitigate against them.

Firstly, the trust of our research participants will be key to the success of this project. The VPN app we design cannot be mistaken by participants for spyware or anything of the sort. The app will, after all, have access to a great deal of their devices' data. But clear data governance policies at our institution are in place to ensure that all data we collect through the app will be stored and processed securely and in line with data stewardship principles. The privacy of research participants will be maintained throughout the data processing lifecycle, from collection and storage, to analysis and deletion.

Secondly, at the communication stage of this project, data shared from the project could potentially triangulate with public data and, therefore, de-anonymise our research participants. This relates with the need to ensure their trust, but also hints at the concern that data is used by third parties in ways they are not intended for. The data gathered through the VPN app is intended solely for the audit of social media platform algorithms. Meanwhile, the code behind the VPN app's architecture

might be used to design VPN services whose intentions are data extraction. Guidance from the VPN Trust Initiative[17] has recently drawn attention to concerns in the VPN industry.[18] Indeed, this study shows that VPN technologies can be used to access great deals of data about how people navigate the web. To avoid any concerns about nefarious reuse of this project's data and code, we will maintain them within the IT infrastructure for one year after the publication of our findings. During this time, the institution's Open Data Committee will be on hand to ensure that those who seek access to the data do so to probe ('reproduce') our findings, or some other acceptable use. The data and code will not leave the environment during that period and will be wiped clean afterwards.

Finally, the risk of data being leaked was mentioned in section 4. Whilst we believe the institution's research ethics policies to be sufficient, we wish to add that the benefit of auditing social media algorithms through real-life usage is far too great to be discarded because of this risk. In other words, the benefits outweigh the risks. Consider that the project will result in analyses of the world's most prominent social media algorithms. In turn, the risk – minimal, as already discussed – is that a few people's online behaviours are captured and they are identified. The benefits are for the many millions of users on those platforms and possibly even future users who gain from the government policies this project informs. The benefits, therefore, outweigh the risks.

17 The VPN Trust Initiative. *The VTI principles*. Available at: https://vpntrust.net/

18 Youngren, J. (2021). 'VPN business ethics research – worrying findings about top VPNs'. *VPNpro*. Available at: https://vpnpro.com/blog/vpn-business-ethics-research-worrying-findings-about-top-vpns/

Appendix 1: Specific prompts for the case studies

Case study 1: Chemical misuse

The below are proposed prompts to aid reflection on Case study 1: Chemical misuse. **It is recommended that the case study is initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some **ethical concerns** raised by this project pertain to:

- anonymisation/pseudonymisation
- what is 'publicly available data'?
- misinformation
- 'vigilante science'
- partnerships
- neuro-linguistic programming (NLP) bias.

Suggested prompts:

- **Anonymisation:** Do Research Ethics Committees (RECS) have standards for anonymisation and pseudonymisation practices? Might anonymising through aggregating data lead to problems of accuracy when it comes to analysis?
- **What is 'publicly available data'?** Are there exemptions for research purposes? Do RECs know or agree about these?
- **'Vigilante science':** Researchers and institutions may adopt a role of 'seeking justice' by aiming at 'pointing fingers' and playing both judge and jury.
- **Partnerships:** What if a project was funded or run in partnership with a government or social media firm? This might add issues such as perverse incentives, biased data, and potential serious real-life harms arising if the results are wrong/misleading.
- **NLP bias issues:** Concerns around bias in the NLP methods, particularly if the data comes from participants who do not speak English.

Case study 2: Synthetic iodide data

The below are proposed prompts to aid reflection on Case study 2: Synthetic iodide data. **It is recommended that the case study is initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some ethical concerns raised by this project pertain to:

- vagueness of methodology
- validity/reproducibility
- technical aptitude of Research Ethics Committee (REC) reviewers
- interoperability
- ethics for different research project formats/methodologies
- environmental impact
- dual-use/unintended use.

Suggested prompts:

- **Vagueness of methodology** is problematic for RECs. Without adequate understanding of the methods a project will employ, RECs cannot identify relevant risks or how projects meet ethics guidelines.
- **Validity and reproducibility:** How do projects employing synthetic data account for accuracy? What challenges do synthetic data pose for the reproducibility of project findings?
- **Technical aptitude of REC reviewers:** What technical skills and understanding do REC reviewers need to adequately engage with cutting-edge data science projects? How are they engaged with ethical questions posed by novel methodologies'?
- **Interoperability:** How do REC reviewers engage with other organisational governance frameworks? For example, the case study states that the 'university has protocols in place for the maintenance of data beyond the completions of such hackathons.' Are these known by the RECs?
- **Ethics for different research projects formats:** Should RECs have different approaches for increasingly diverse projects? This case is a hackathon, which is not a standard research format. What about workshops? Should all research be scrutinised by RECs?

Case study 3: Spiking Neural Networks

The below are proposed prompts to aid reflection on Case study 3: Spiking Neural Networks. **It is recommended that the case study is initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some ethical concerns raised by this project pertain to:

- intellectual property (IP)
- bioethics
- dual-use
- deception
- payment.

Suggested prompts:

- **Intellectual property (IP):** Who owns the data of the Myo sensor? What if the data is stored and accessed through a third-party server, in particular run by a commercial entity? What if the study is run in partnership with a private company?
- **Bioethics:** Narratives and models of AI are informed by biological sciences. AI might be claimed to emulate the nervous system and brain functions. The same models might then be used to inform future research into the cognitive sciences. What happens if these assumptions are wrong, since there is a lack of understanding of the complexity of the brain?
- **Dual-use:** Spiking Neural Networks (SNNs) are powerful neural networks with great processing capacity. What if these are increasingly used to create increasingly powerful AI systems? For example, could SNNs be used to teach a robot hand to improve the technique of throwing a hand grenade or maximise efficiency in automated work contexts? What if the company mainly produces military-grade applications? How do we ensure the development of SNNs such that the technology only apply to societally beneficial uses?
- **Deception:** What if participants are asked to perform unexpected tasks that they have not consented to and that they find morally problematic?
- **Payment:** How do we ensure participants are adequately reimbursed for their contribution to the research? How might they be acknowledged?

Case study 4: Tattoo ID to map gangs

The below are proposed prompts to aid reflection on Case study 4: Tattoo ID to map gangs. **It is recommended that the case study is initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some ethical concerns raised by this project pertain to:

- data provenance
- consent
- privacy
- social injustice
- unintended consequences
- data security.

Suggested prompts:

- **Data provenance:** How reliable are the data sources? Do tattoos reflect a gang's location, name, and even specific information about the role the individual plays in the gang? How is the accuracy of image annotation ensured? How might this affect the efficacy of the trained model?
- **Consent**: Have people been informed during arrests or pre-incarceration inmate interviews that the images of their tattoos will be used to create a database to train an AI model? How about consent for the use of images taken during surveillance operations?
- **Privacy:** Can people be identified based on the images of their tattoos, even if their names are not stored? Can the data be cross-referenced with other databases to infer identity?
- **Social injustice:** Could this image categorisation system be used for racial profiling? What unjust assumptions can shape such an analysis?
- **Unintended consequences:** What happens if people are misclassified and sent to a penitentiary with rival gangs present? How are humans kept in the loop to avoid such mistakes?
- **Data security:** Who is authorised to access the data? Can the data be used by other government agencies in the future, and what conditions must be met for data to be shared?

Case study 5: Emotion recognition

The below are proposed prompts to aid reflection on Case study 5: Emotion recognition. **It is recommended that the case study be initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some ethical concerns raised by this project pertain to:

- data labelling
- privacy and consent
- stakeholder engagement
- dual-use
- partnerships
- deception and harm.

Suggested prompts:

- **Data labelling:**
  - For supervised learning, the datasets require annotators to use basic emotion models for labelling the data. Are these models adequate to understand affective states? Can emotions be categorised in this way? How might annotators be open to bias when labelling images or video content? What skills must annotators have?
  - For unsupervised learning models, can emotions be recognised from static facial expressions identified through web queries, or on social media platforms? What contextual information might help identify an emotion?
- **Consent:** Have people given their consent for their images to be labelled for the purpose of training the AI model? Who owns the data available through websites? Is there some expectation of privacy when personal images are posted online?
- **Stakeholder engagement:** How does the project engage with autistic people? Is this engagement sufficient to ensure the developed technologies respond to their needs? How might a more participatory approach improve the project's outputs?
- **Dual-use:** The research output could be used to inform facial recognition technologies in the healthcare context, e.g. monitoring of compliance with treatment, diagnosis of mental health conditions etc. What impact could this have on patients?

- **Partnerships:** Is there a possibility that this software becomes commercialised? How would that affect your reasoning about its ethics?
- **Deception and harm:** Potentially negative emotions could be caused in the participants through recognising emotions and showing emotions to the participants that are not necessarily accurate.

Case study 6: VPN app

The below are proposed prompts to aid reflection on Case study 6: VPN app. **It is recommended that the case study is initially reviewed only with the general prompt (provided on the cover sheet) in mind.**

Some ethical concerns raised by this project pertain to:

- timeliness
- data deluge
- manipulation
- lack of awareness
- self-evaluation.

Suggested prompts:

- **Timeliness:** The research proposal states that the VPN app that will be used to gather data about online algorithms has already been designed and is at its final development phase. Should the proposal have reached the Research Ethics Committee (REC) sooner? What risks could the VPN app pose that would have otherwise been mitigated against had this been reviewed by the REC sooner?
- **Data deluge:** The VPN collects data from across any one device rather than collecting targeted data from select social media platforms. The proposal mentions GDPR's principle of 'data minimisation' but fails to explain how data is minimised other than having an algorithm sift through the collected data, and a researcher deletes unnecessary data once it reaches the institution's servers. Are these measures sufficient? What guidelines might the researcher follow to ensure only relevant data are stored for analysis? Might too much data still reach the analysts, and what concerns does this raise?

- **Manipulation:** The app research participants will install on their devices is referred to as a 'VPN'. Whilst the proposal claims it will have the basic functions of a VPN, this still seems to be a way to keep participants from seeing the app as deeply invasive. As mentioned in the 'data deluge' concern, a great deal of information is being collected from these individuals, and these risks are downplayed by naming the app a 'VPN'.
- **Lack of awareness:** The personal nature of the data compiled is not adequately discussed in the proposal. Do the researchers understand the intrusiveness of the app they are having research participants installed? And is the 'VPN app' the best way to go about researching social media algorithms?
- **Self-evaluation:** The final section proposes a cost-benefit analysis, where the cost is the risk posed to the research participants, and the benefit is the use of auditing social media algorithms. How might this analysis be improved? What does it say about the REC process when researchers give themselves ethical approval?

Preferred citation: Ada Lovelace Institute. (2022). *Looking before we leap: Case studies.* Available at: https://www.adalovelaceinstitute.org/resource/research-ethics-case-studies/