Please send comments to the special interest group committee data-ethics@rss.org.uk

Reviewing the data ethics landscape - addressing some questions

1 Introduction and Summary

The new data ethics special interest group (SIG) of the Royal Statistical Society (RSS) has a remit across the breadth of work with data. This includes many different types of data, and existing practices as well as diverse ethical implications which are already being discussed. Thus, the group has reviewed the landscape to look at what is being done, and address four specific issues of what to do, who to work with, professional and public implications. Following this introduction and summary, section 2 clarifies what we mean about practical ethics, and more specifically what data is used for. Sections 3 to 6 cover established meta-ethical concerns of autonomy, justice, public benefit and privacy, each having new developments relating to data. Section 7 draws out five normative ethical challenges emerging specific to data which demand further attention which the RSS would like to see action on. Section 8 draws a more strategic imperative for the RSS, as a gap in what is described as data governance, i.e. the accountability for organising the ethical use of data. The largest gap in understanding is the international landscape where we begin to identify some issues based on other sources and second-hand information (section 9). We conclude with some proposals of initial areas of practical focus for the SIG in section 10. A short list of activity in the UK is included as an appendix.

The paper can be summarised via the following four sets of questions:

1.1 What are the gaps? Where should the RSS focus?

Data ethics may be understood via three ethical registers: meta-ethics which asks the big questions about what is right and wrong, normative ethics which asks what ought to be done, and applied ethics which asks what is, can or must be done. In contrast to the practicalities, the more abstract ideas of ethical principles, norms and mores are where others will be expert. But each will bring different considerations of data ethics in the context of people’s rights and professionals’ and researchers’ responsibilities. In practical terms, the nature of data and how it is used must be understood but as yet neither of these complexities are being embraced effectively. Concepts of fairness and justice apply to whole populations and relative outcomes yet typical analysis focuses on absolutes, a sample or one group as a contrast rather than variegations in the whole population. Outdated consent and privacy principles dominate discussions but are mismatched to the nature and use of big data. Societal benefit contrasts with individualised harms, but the former is harder to define, measure and articulate. While the RSS can criticise the suitability of principles for practice, its main role is in proposing better solutions or necessary compromises. Emerging issues around the full data landscape’s complexity and its governance warrant attention are where we propose to foucs.

1.2 Who should we work with externally? How does this fit with existing RSS plans?

Internationally, we can identify and contrast our strengths and work to share practice, relying on Ada Lovelace Institute and UK Statistics Authority as signal bodies. There are many who already work on ethics in health data and we should collaborate where we can add value, perhaps in responded to consultations and inviting speakers and attending external events. While people are working on social media data in limited ways, on geospatial there are no obvious partners, although the ODI have identified the benefits flowing from sharing data and therefore potentially foregone by failure. Other emerging issues from codes and regulations to time and capabilities could be put to newly established bodies for such advisory and deliberative purposes. Defining concepts is best done in an academic context in conference and journal submissions, not least through the RSS. Challenging existing professional practice and setting appropriate standards is work developing in Data Science Section[1], and by convening practitioners to discuss and establish norms. Other Sections have meeting which we can co-sponsor and attend, including Statistics and Law, Official Statistics and Emerging Applications. We should also represent professionally on oversight boards and encourage others to do the same as the demand is great and our number few.

1.3 How should we develop the training offered? What are the professional implications?
Ethical review in research practice is well established and the norms enshrined therein should be applied to data. Tools and associated training have been created by bodies with substantial interest in data and specialised ethics boards are emerging to suit different functions. Professionals accustomed to primary data being collected for analysis by direct contact with research participants will be familiar with team approaches and approvals processes. These processes themselves are not designed for use of already-collected data, and the internal use of data collected for service delivery arenas lack governance exemplars. Training about the application of existing ethical practices to extant data is needed at the level of ethics bodies (through UKRIO), project management (particularly internal projects), as well as updating what is offered to new professionals. The Data Science Section is also looking at ethical principles within an agile workflow, and other curriculum initiatives have included ethics (RSS members are involved). However, what is lacking is the nature of curriculum, and the way of embedding this into existing teaching in order to change norms in practice. While case studies are developing, there can be a role for the RSS to develop exemplar training materials or develop a model to ‘train the trainers.

1.4 How is the media informing the public? How can we know the public view?
Media reports about artificial intelligence are speculative and futuristic rather than about practicalities or focused on ‘privacy’ of large databases and individualised surveillance. However, opt-in/opt-out choice defaults, automated decision-making and online targeting have raised statistical implications, and more recent reports are providing some balance on benefits. Media reporting of data stories is limited compared to popular topics like health, and so most reporting is health applications, but all reporting has implicit challenges to the narrative of social licence to use data. Health researchers have been most developed in their public engagement, from a tradition of PPI and ethical requirements for consent and information. Citizen juries are finding the public has nuanced views on what is appropriate and want to be consulted and respected by processes but they have very little understanding of big data. Broader public opinion research is nascent and general survey results undermined by limited understanding although Which? has identified categories of online data sharing behaviour. The RSS can engage a range of groups to understand advocacy for

[1] They have a paper on professionalisation and industrialisation based on workshops with data scientists https://www.rss.org.uk/Images/PDF/get-involved/2018/Professionalisation%20of%20DS%20-%20summary%20of%20workshop%20outputs%20for%20publication.pdf

different positions on data sharing, and their basis, the SIG will hold meetings and encourage research on public engagement.

2 Defining Ethics and Big Data Usage

Data arises in new situations, and it gives responsibilities to people who would not have had ethical discretion previously[2]. Ethics is about both agreeing what ought to be done and doing it - norms in practice[3]. While the applied ethics of practice is particularly relevant to statisticians who are likely to be responsible in the new situations, statistical expertise is also important in informing what the new situations are and what ought to be done[4]. A normative ethics which accounts for the situatedness of data production and analysis would imply proscriptive precautions should be based on assessment of risks. The RSS should challenge the overtly individualistic ethics being used in restrictive compliance models and promote relational, situational and virtue ethics - thinking about what we ought to do.

The nature of 'data' is essential to understanding the ethical challenges it presents. The word has historically been given plural usage from its derivation as 'givens'. This usage is increasingly unsuitable as data is not readable lines in a database, and those categories which are in the data are not given but socially constructed. More generally data in use comprises the stochastic idea of how certain matches are in linked data, and data which indicates that records are missing. Data is an instantiation of an information space, coded according to conventions, with various sources of error. Thus 'data' is used here as an abstract noun which is grammatically singular (the RSS could get this across clearly and consistently).

Emergent as a problem is that disciplines with strong ethical traditions have not managed to grasp the nature of big data[5]. Such framing as a 'big database' is typical of public understanding of administrative data, but bioethicists are confident saying big data is not well defined.[6] The recourse to descriptions of velocity, variety, volume, veracity and complexity miss the point about how this changes the workflow of the analyst. Small data is directly auditable by structuring into a table which could be reviewed if not entirely, systematically: all the big data descriptors preclude that. The data science workflow has meant statisticians receive data without traditional gatekeepers, and interactions with the data are entirely mediated by computational interfaces. The RSS needs to engage with the public and regulatory understanding of big data and make sure the nature of its use is conceptually understood, drawing on serious practical examples.

Data is accumulated as part of a primary activity but can then be stored for other uses. The separation of uses beyond direct service provision to planning, research and commercialisation has

[2] Tractenberg, RE "Institutionalizing Ethical Reasoning: Integrating the ASA's Ethical Guidelines for Professional Practice into Course, Program, and Curriculum" pp.115-139 in Ethical Reasoning for Big Data

[3] Nuffield Council on Bioethics, The collection, linking and use of data in biomedical research and health care: ethical issues, London: Nuffield Foundation, 2015

[4] Ethical Reasoning for Big Data, Collmann & Matei (Eds.), Basel, CH: Springer, 2016

[5] UKRIO notes universities have limited ethical scrutiny of big data research applications

[6] Ethics of Biomedical Big Data, Mittelstadt & Floridi (Eds.), Basel, CH: Springer, 2016

been made in health.[7] Further categories, of statistics and management information, might also be added - the data is both part of the service and part of its governance.[8] The categorisation of the big data 4Rs: reuse, repurposing, recombining, reanalysing[9] highlight that data is used in many ways beyond its origin. At each stage boundaries can be blurred, as data is used to validate other processes, each of which may include errors. Commercial contracts and external research applications generally have a clear approval process and restrictions on usage. However, within compliance with regulations, ethics needs governance and internal usage should have regard to ethical frameworks. The RSS can promote documentation of uses and workflows, so that principles and their ethical basis is clear at each stage and ethics can be embedded within training[10].

3 Fairness and Justice

Data in itself comprises social judgements: what is important to record, and how to record that. Use of data then describes associations between these judgements, and social discrimination becomes embedded in the empirical[11]. While there are extant professional imperatives to report the statistical limitations of the inference[12], sensitive cultural conclusions need attention. Specifically, the Expert Group of the European Data Protection Supervisor identifies respect for the dignity of groups as a new ethical consideration[13]. Socially pejorative conclusions are readily implied in the reporting of associations of groups to circumstances without suitable insight into causal mechanisms. Statisticians know that correlation does not demonstrate causation, but in the absence of alternative explanations it is a compelling public narrative. The RSS can recognise that professional practice does not stop at estimating a correlation and walking away from the implications by saying the inference is not causal.

Coverage of data includes what is measured, and who or what it was measured from[14]. Which units are missing is of particular importance when a group targeted for policy intervention has lower representation in the data. Limited data also increases uncertainty in inferences, leading to more conservative conclusions for groups with sparser coverage. More than this, a sample of data is representative of those it was collected from and excluded populations may be very poorly served. These are all instances of fairness which relate to the coverage of the data, most obviously consequential in training data for algorithms. However, this issue is well documented by many authors and may be an area for collaboration, with some more attention given to conceptualisation of original measurement categories. The RSS can highlight that inferences based on a subset of a

[7] Review of Data Security, Consent and Opt-Outs, National Data Guardian for Health and Care, Leeds: DoHSC, 2016 www.gov.uk/government/uploads/system/uploads/attachment_data/file/535024/data-security-review.PDF

[8] Transformation guidance for audit committees, National Audit Office, London: TSO, 2018

[9] Ethical Reasoning for Big Data

[10] In his presidential valediction, David Spiegelhalter anticipated the ethics would cease to be isolated, p.2 in https://www.rss.org.uk/Images/PDF/about/2018/RSS-Annual-Report-2018.pdf

[11] Broad, E, Made by Humans, Melbourne, AUS: Melbourne University Press, 2018.

[12] Tractenberg

[13] Ethics Advisory Group, Towards a digital ethics, European Data Protection Supervisor, Brussels: European Commission, 2018

[14] Economic and Social Science Research Council, Longitudinal Studies Strategic Review 2017, Swindon: UKRI, 2018

social group may not be appropriate for use in a broader social policy context and such extrapolations diminished.

A further instance of fairness relates to discrimination: certain ‘protected’ characteristics are proscribed from featuring in the reasoning for official decisions. But the purpose of much use of data is literally to discriminate, to allocate subjects to groups - an allocation can always appear unreasonable to those not favoured and universal fairness is mathematically impossible[15]. Where automated allocation is trained by existing allocations, it will aim to replicate the optimum of the status quo seen in the training data. While this may cause society to review its own fairness, it is incoherent to expect trained automation to transcend the human biases it is based on, even as we recognise a societal problem. Applied ethics demands a practical response, such as a comparison of new allocations to old, and looking to improve against agreed criteria as objectives for a new project. The 2018 Budget announced that the newly-established Centre for Data Ethics and Innovation (CDEI) was going to review bias and micro-targeting for online decision making, in line with a recommendation from Which?[16]. Issues of coverage are now being taken seriously so it is not clear that the RSS has anything further to add beyond critiquing tools used in the assessment of fairness.

4 Freedom and Respect

Respect for the agency of individuals, as a human rights concern, is usually achieved through specific, informed and personal consent. Data reuse has excluded each of these three aspects of consent, through vague and general terms of service, and using indirect, regulatory routes to approve use of existing data[17]. The value of large data is that it already exists, so having to contact every person would compromise its utility and respect for persons must be achieved in other ways: by producing suitable information and seeking people’s views about uses, and governance[18]. This respect goes as far as facilitating meaningful objections to using data, at any stage in the work, balancing powerful demands, public values and minority views.[19] The governance problem is an area of special interest to the RSS as it delimits the use and curation of data, as well as relations between organisations.

The language used to describe data, how it is used, and its governance is technical, making it difficult for consent to be informed. Many terms have both technical and social meanings in this context. Trust can refer to a corporate entity responsible for data stewardship; confidence can refer to the common law duty. ‘Planning’ is used in the health context to describe use of data not for direct delivery of a service but not for research. Members of the public not familiar with these terms may not interpret them in the ways that professionals do - clumping planning with research is

[15] Dwork, C has various work on composition effects in fair processes - it is complex

[16] Britain Thinks, Ctrl, Alt or Delete? Consumer research on attitudes to data collection and use, London: Which? 2018

[17] MRC guidance on GDPR, Using Information about People in Health Research, Swindon: UKRI, 2018

[18] Elias, P, Research ethics and new forms of data for social and economic research, Paris: OECD Publishing, 2016

[19] Sunstein, C, The ethics of influence, Cambridge: CUP, 2016

proposed because planning is not understood, conflating internal and external uses[20]. The deployment of a number of terms should adapt to allow for what is understood by the public rather than the technical expectation. The RSS might include this concern in training for its ambassadors etc or recommend guidance on plain language aimed at the public, drawing on insights from ESRC and others.

Public opinion on data sharing has been sought for some time, but it readily became clear that the public is unfamiliar with administrative data[21]. Moreover, the simple idea of whether something is 'trusted' is far away from the detail needed to understand public views. This presents radical challenges to the requirements currently on those who work with data: they should both explain their work and consult with the public about its legitimacy[22]. Medical researchers have been most bold in this area, developing requirements for publicising data and its usage, and in using a range of approaches to working with in publics[23], for example citizen juries, deliberative workshops, ECOUTER, World Cafe and focus groups. Interesting results are emerging such as patients being very much more positive about data use (if unrealistic)[24] but such endorsements should take into account the vulnerability of their health status[25] (and moderate their expectations). Surveying public views on heterogeneity of values, interest and understanding is naturally a subject of interest to RSS members, especially for public data and data policy.

5 Public Benefits

Some radical reconsiderations arise, taking us away from control over data legitimated by its initial collection, in a paternal model. This 'social licence' to determine what benefits the public is exclusive, elitist and entitled, none of which attitudes respect the people the data come from[26]. Reorienting ethical considerations to the use of data to benefit people, and account to them on the benefits should find quite different and diverse priorities. The RSS may wish to examine the nature of public benefits from data use and sharing in contrast to overblown promises, about which the public are rather sceptical. The RSS should also show leadership in making use of data accountable to people and presenting the case for data use in public forums, allowing the public to arbitrate what their good is.

While funders of health research in the UK require appropriate public and patient engagement, disciplinary constraints on use are also introduced by professionals. The restriction in health applications sees specifically medical outcomes legitimated because doctors (typically GPs)

[20] Data sharing and the importance of choice architecture in healthcare: new results, Behavioural Insights Team, Research for National Data Opt-out Programme https://www.behaviouralinsights.co.uk/trial-results/data-sharing-and-the-importance-of-choice-architecture-in-healthcare-new-results/

[21] Cameron, D, Pope, S & Clemence, M, Exploring the public's views on using administrative data for research purposes, Ipsos Mori, Swindon: ONS & ESRC, 2014

[22] Sunstein

[23] Thorogood A, et al. APPLaUD: access for patients and participants to individual level uninterpreted genomic data. Human genomics. 2018, 12(1):7. Kaye J, et al. Including all voices in international data-sharing governance. Human genomics,12(1):13, 2016

[24] useMYdata www.usemydata.org/casestudies.shtml

[25] Ethics of biomedical big data

[26] Carter, P, Laurie, GT & Dixon-Woods, M, The social licence for research: why care.data ran into trouble, Journal of Medical Ethics, 41:404-409, 2016

controlled the initial patient data collection; health bodies further restrict data use, both without reference to public expectations. Lay involvement to determine uses of data for the public good is overlooked in legal structures to protect data conservatively, without considering ethical imperatives to share. Similarly, the lay public sincerely doubt the ability of good theoretical ideas to realise beneficial impacts[27], even as the imperative for research to have impact is influencing the projects proposed. Good practice in making the public case for the use of data, in ways that are comprehensible, need development but the examples of plain language summaries and participant engagement in METADAC are indicative.[28]

The term ‘trustworthiness’ is difficult as although its aspiration is coherent, there is still the implication of an appeal to an elite judgement on its character. This is more important where consent and public autonomy relate to broadly drawn uses of data whether through ‘broad consent’, commercial terms of service or the implicit consent of public service provision. Again this is problem of governance that there can be a public expectation of behaviour which governance should elicit and respect. Conversely usage should be transparent and possible for public to challenge. Models for governance in this way are nascent[29] and an area where the RSS could lead given the nexus of expertise on data and engagement of producers.

Media reporting of the ethical basis for use of data has had the hysterical features of much mass journalism balanced by some countervailing expertise on what is actually happening[30]. This is deficient in two ways: the actual working of emergent data usage is complex[31]; and information about data collection and usage is obfuscated by government and businesses[32]. Scandal can then be attached to usage which is misunderstood[33], and trivialities for which appropriate permissions were never sought. A clearer picture of actual possibilities rather than the obsessive focus on health and robotics would be more helpful and the RSS could develop more case studies based on high quality examples[34]. The RSS could offer some guidance about media reporting about AI, or engage with organisations such as the Science Media Centre and broadcasters such as the BBC to develop broader and critical reporting.

6 Reconceptualising Privacy

[27] Tully, MP et al., Investigating the Extent to Which Patients Should Control Access to Patient Records for Research: A Deliberative Process Using Citizens' Juries, Journal of Medical Internet Research, 20(3):e112, 2018

[28] Murtagh, MJ et al. Better governance, better access: practising responsible data sharing in the METADAC governance infrastructure, Human Genomics, 12:24, 2018

[29] METADAC is one of the most advanced but quite constrained; similarly NS-DEC

[30] Longer articles are appearing which offer more depth and balance including in MailOnline from 06/12/2018: https://www.dailymail.co.uk/health/article-6464033/Thousands-people-rare-diseases-DNA-recording-study.html

[31] Collmann & Matei identify reuse, repurposing, reanalysis and recombining in preference to focus on size

[32] Business scandals abound but government data collection from businesses is reviewed in Bulk Collection, Cate & Dempsey (Eds.), New York, NY: OUP, 2018

[33] A good example is the data science work of the Behavioural Insights Team, Using Data Science in Policy, London: Cabinet Office, 2017

[34] e.g. Big Data: does size matter?, Timandra Harkness, London: Bloomsbury, 2016

Personal privacy for abstract data is something new beyond protection against identification in anonymous data, coupled with the novelty of digital identity[35]. This involves the sensitivity of inferential disclosure of new information in the context of what was already known. More generally, the sensitivity of information recorded is understood in relation to the other data in context. Thus linking data together can increase its sensitivity e.g. linking a historic indiscretion to a contemporary public persona. However, public privacy demands have been taken to correspond to strict confidentiality or anonymity which are neither feasible, nor societally expected[36]. Some progress is being made on what sort of privacy is expected but it is easy to ask questions which misspecify public concern[37].

Ethical considerations most obviously occur in personal data, but most new data relates to people, in how services are provided to people, or used by them. These arise a specific new challenge, in that the transactions happening are social facts, known to anyone who observes them.[38] Your neighbour knows where you live, without any regard to your consent your movements will be observed by those who see you. Every person has a geospatial location and recording this systematically is not a problem in itself in isolation. But these digital traces build up to bigger pictures which can be disconcerting as we do not realise we leave them, particularly if they can be linked to reveal information that people would rather be kept private. It is not realistic to obliterate such traces, not least as they would restrict popular uses such as efficient transport services[39]. Regulating their usage is however more feasible so the RSS SIG has specifically sought expertise on geospatial data ethics.

7 Emergent Normative Challenges

Some of the challenges emerging in practical terms are not areas where the RSS has expertise, but could collaborate with others.

7.1 Time

Temporality presents us a number of challenges. It is an important component of metadata, and offers structure to linked data, so enhances its value. The time of collection embeds also the data standards at that time, such as measurement and classification. But also includes the social and cultural understanding of what it is both possible and appropriate to use data for. Conversely, uses of data take time and can therefore experience societal changes and need the governance structures to respond to them. There is also an awareness that machine learning once trained can become out of date, or if set to evolve can create problems. Governance may help in some ways but good practice guidance about review and audit are more likely to be context specific. The RSS could consider this within the working party Data Science Section are running with the Institute and Faculty of Actuaries.

[35] Cate & Dempsey

[36] The Politics of Big Data: big data, big brother? Saetnan, Schneider & Green (Eds.), London: Routledge, 2018

[37] Ethics of biomedical big data

[38] Seven veils of privacy are proposed by O'Hara, K, Privacy: Essentially Contested, a Family Resemblance Concept, or a Family of Conceptions? Presented at Amsterdam Privacy Conference, 2018

[39] Open Data Institute, Personal data in transport: exploring a framework for the future, London: ODI, 2018

7.2 Skills

Skills for ethical use of data extend from being able to make ethical decisions in practice, to being able to engage as a citizen. In a structured way people within organisations need to understand the ethical implications of their work with data, and toolkits have emerged for this purpose. Where this gets more complex is the need to engage with external organisations and authorities e.g. using third party data, and case study training approaches have been simple so far. The ASA Committee on Professional Ethics has developed some work with an indication of low uptake and poor learning under current models[40]. Data governance as a concept is still emergent to deal with this, but if it includes lay perspectives directly it will deal with the other difficulty: Effective communication with the subjects of the data. The RSS should consider what training is needed and how it can contribute through professional courses and other offerings but ethicists also need to learn about statistics[41]. The most appealing delivery model is to meet new demand in Data Science (and related) MSc programmes through a 'train the trainers' project[42]. This would require considerable innovation so the RSS would need to identify suitable partners and funding, aiming to embed in the curriculum.

7.3 Ethical Codes

Professionals have developed codes for their professional practice, under a model of self-regulation, proliferating now in data ethics[43]. This is effective in areas where the public is best served by work to a high professional standard, as judged by peers. Ethical use of data is more likely to be better considered through a model of co-regulation, with external advisory input. While professional codes should consider issues of data ethics, the essential nature of real data to professional practice means the issue has been neglected[44]. Sensitive regulatory intervention, to recommend, convene, review on the content of codes should be encouraged. More support may be needed in organisations which are small or disparate, and need an external ethics review group to provide advice[45]. Putting professional codes into practice is an important problem which the RSS might recommend others attend to through various models of co-regulation[46] and also integrate into its own professional code of conduct.

7.4 Regulation

Regulations for data are already in place, not least through GDPR[47]. The statutory basis and powers on recent bodies such as the National Data Guardian and Centre for Data Ethics and Innovation and how they demarcate the landscape with regard to data uses as well as established bodies such as the Information Commissioner is unclear. Legal redress for GDPR breaches requires a claim of harm to be made which affected an individual, putting ownership responsibilities on organisations. While

[40] Tractenberg

[41] https://www.rss.org.uk/Images/PDF/about/2018/RSS-Annual-Report-2018.pdf

[42] The ESRC RMTC call lists this as a potential model of training and would host materials (p.5) esrc.ukri.org/files/funding/funding-opportunities/research-methods-training-centre-call-specification/

[43] Floridi, L et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds & Machines, 28: 689, 2018

[44] Tractenberg

[45] An example is the Machine Intelligence Garage www.migarage.ai/ethics-committee/

[46] Cave, J, Marsden, C & Simmons, S, Options for and effectiveness of internet self- and co-regulation, RAND, Brussels: European Commission, 2008

[47] Hand, D, Aspects of Data Ethics in a Changing World: Where Are We Now? Big Data, 6(3):176-190, 2018

the consideration of natural persons is essential to the legal system, it forces focus on deterministic and tractable conclusions. This makes concerns for groups, and uncertainty about individual much harder to judge, as is the (monopolistic) control of data[48]. For example, has private information which is estimated with a probability of 10% to be correct been negligently disclosed if it is found to be true and a harm occurs? While there is currently review of the implementation of GDPR, the RSS plans a meeting with Social Statistics Section looking at how it is distorting thinking.

7.5 Principles

Existing ethical principles from bioethics are based around the individual as much as society: beneficence, non-maleficence, justice, agency[49]. These make it very difficult to discuss the societal benefits of data sharing which improves efficiency of services or enhances government. The attempt to require public good tends to excludes the commercial utility of enterprise which benefit society, even as other frameworks are available[50]. Even so, who is then called upon to adjudge the public good of a project is not clear - the public themselves may be more permissive about some uses and more restrictive about others than official bodies. New ethical terms such as 'solidarity' i.e. benefits due to a community or future populations have recently emerged[51] but not been universally embraced. Others such as 'explainability'[52] may be better understood as principles of ethical governance than new ethical principles. Most likely the RSS can organise around putting existing ethical principles into the practice of big data research and contrasting the theoretical properties of new ones with practice.

8 Governing the Sharing of Data

It is desirable to share data, to improve provisions as well as making them[53]. So privacy facilitated by data ownership requires strong contractual relations which are at odds with the market positions of consumers. Rights derived from ethical principles are an abstract solution - we often hear of human rights violations but need structures which uphold them. The ideas that remain are about data stewardship and data governance, both of which rely on people to form judgements about the appropriate thing to do, in relation to other people. Every person bears moral responsibility for the ethical use of data at every stage but accountability may be structured to confer legal expectations on some and not others. The dismal failure to conceptualise the social nature of governance and the involvement of individual people in decisions about it is an area the SIG should attend to directly.

Stewardship is an incongruous term in the context of the proprietary models for data which privilege initial aggregation as ownership over sharing. But the social basis of sharing is more appropriate in that it is a social contract not to discriminate, to impose appropriate expectations on others, and attest to the provenance of the data. Moreover, the social licence to share requires an understanding of mutual dependency between those who provide their personal data and those who

[48] Shah, H, Use our personal data for the common good, Nature 556, 7, 2018

[49] The Menlo Report: Ethical principles guiding information and communication technology research, Washington, DC: DHS, 2012

[50] Dignity, autonomy, welfare and self-governance are offered in Sunstein

[51] Prainsack, B & Buyx, A, Solidarity in biomedicine and beyond, Cambridge: CUP, 2017

[52] AI in the UK: ready, willing and able? House of Lords Select Committee on Artificial Intelligence, HL100, London: TSO, 2018

[53] Open Data Institute

process it. Specifically, the use of data can realise a public benefit which can justify its use and encourage people to give their data. But how a public benefit is seen to justify one use and not another requires further exploration, not least what the accountability is for this justification. For example, the exclusion of commercial use of health data is conservative rather than ethical in basis, precluding public benefits arising from commercial uses[54]. The RSS might promote decisions based on ethical principles, and transition to such positions facilitated by public debate, say through the Ada Lovelace Institute.

Competing interests are often a concern in governance roles, but in the context of data, every person has an interest in features of their own. So long as there is diversity among those in governance roles, these interests can be balanced but more powerful interests may still dominate. More difficult is the need for expertise in the nature of data and its usage, which typically come from professional experience, bringing other interests. As all actors from companies to government are struggling with the ethical implications of data, a balance of interests openly declared may yet be effective, if there are people to criticise the balance. A shortage of people with expertise on data ethics requires specification of the balance and its interdisciplinary nature[55]. The RSS might also wish to identify shortages among people to fulfill governance roles, especially as governance roles may relate to each other and therefore preclude individuals from multiple appointments.

Governance should be based on people who understand the issues and how they relate to people who have different interests. The aim is to generate trustworthiness by facilitating consideration of concerns and having a rational basis for decisions, without explicitly, or by default, favouring powerful groups. However, trustworthiness is an aspiration which needs to be based on practical actions, not a brand (similarly 'safe'), so ethical issues are known to have been capably considered. Communicating openly at every stage, so that actions and rationales can be debated and contested will require public explanations, not just technical reports and meeting minutes. The RSS has both the independence and the authority to call for comprehensible, open and accountable publication of the business of these governance structures.[56]

9 International Practice

Data is not constrained by national boundaries: It links between points conceptually without reference to ownership. It can exist in multiple places at once and also characterise international activity[57]. So international approaches are important, both collaborating and leading. A number of countries say they aim to lead on the ethics of AI - trust will be a commodity which enhance a national brand. But on the less glamorous but more pervasive ethics of data, there is silence. The UK has considerable presence in an international landscape that is poorly understood. Large public data systems in the health sector have struggled through crises of permissiveness, and against presumptive social licence, and now against restrictive compliance. Similarly the UK official statistics function has access to the largest extent of data for government statisticians anywhere.

[54] advertising type use excluded options are kept open in Academy of Medical Sciences, Our data-driven future in healthcare: People and partnerships at the heart of health related technologies, London: AMS, 2018

[55] Both CDEI and Ada did this but Murtagh et al. offers an exemplar development in practice

[56] Spiegelhalter, D, Trust in numbers, Journal of the Royal Statistical Society, Series A (Statistics in Society), 180(4):949-965, 2017

[57] Cate & Dempsey

Both have evolved several generations of governance for use of public data where other countries still believe in their social licence which is largely social ignorance.

Like a number of other countries, the UK has set out an aspiration to lead the world in the ethics of new uses of data. There is no comprehensive review available to say whether this is the case, but many other countries are establishing ethical reviews and responsible bodies. What we can find is that bodies in other countries are focused on more advanced work such as algorithms (in New Zealand) or AI (in Germany). Many countries also have a privacy commission of some kind[58], which is separate to the statistics institute. We have encouraged the CDEI to understand the international landscape[59] and we hope it will make this a priority which involves the RSS Data Science Section. In complement, the Data Ethics SIG can try to establish what other practice is emerging in other countries, in the context of differing capacity and structures.

The RSS Section meeting in June, ‘Analysing without Consent’[60] was distinctive in several ways. It was not about machine learning, algorithms or AI, it was just about data. It covered various disciplines and origins of data and governance structures[61] rather than focusing in one domain. It was focused on practice i.e. what one could do in the light of an ethical challenge. As such the meeting demonstrated that the UK is leading in this particular area and so should be encouraged to host work on fundamental of data ethics. The UKSA is planning an international conference in 2019 and the RSS should seek not only involvement but to lead additional and complementary programme. It also seems logical that the SIG should continue its work to review this area and publish a strategic review of the fundamental data ethics landscape, with particular attention to governance and the emerging features described earlier.

10 Domains of Potential

Effectiveness of public engagement needs to develop but health, social media and geospatial are specific data which warrant particular attention by the RSS, for slightly different reasons:

- Health has a lot of data and systems which support its use in practical applications, but it also has substantial regulation, public concern, and real risks. Partnerships and expert comment are important, not least to bridge wider use of data for human flourishing in their lives outside clinical settings. Keeping up with the policy discourse will need substantial commitment from RSS members to lead this work and engage with others in developing good practice.
- Social media data (and more generally Internet tracking data) are already being applied without individual awareness let alone deliberative societal consent. There are questions of what is possible, and the impact use of these data has, as well as meaningful input by the public, but strategic convening is currently lacking. Data Science Section may wish to take an interest in how online data is used and what is appropriate, and the ethical implications.

[58] Debating Ethics, 40th Annual Conference of Data Protection and Privacy Commissioners, hosted by the European Commission in 2018 attracted more than 100 such www.privacyconference2018.org/en

[59] https://www.rss.org.uk/Images/PDF/influencing-change/2018/RSS-CDEI-submission-5-Sept2018.pdf

[60] www.statslife.org.uk/members-area/sections-and-local-group-meeting-reports/3830-emerging-applications-and-social-statistics-section-meeting-analysing-without-consent

[61] This was apparent as the speakers typically did not know many of each other

- Geospatial data is fundamental to the delivery of services, particularly public services, as services for people have to be delivered to them. Structuring delivery of services for and by people requires understanding of location and nature of demand and infrastructure to deliver. Moreover, every person has these transactions in locations, and the data associated is fundamental to realising their needs but what are the ethical concerns to be negotiated? This seems to be an area where much more thinking and partnerships are needed and one in which the RSS can only play a part, so it needs to raise awareness of concerns.
- Public engagement may provide an antidote to paternalistic presumptions of social licence. It also challenges zeal in regulation and privacy groups - where data users feel systems are not proportionate, reference to public views may resolve this. Most people, including ethicists, have no idea what big data looks like let alone how uses can impact society. Engagement cannot be effective without taking on these concerns, suggesting new interdisciplinary work, engagement activities and survey question development from RSS members to describe public opinion, including opportunities for co-production of data ethics and data governance practices.

There will be other areas to attend to, particularly resolving the issue of data ethics training and the other emergent issues of section 7 but this is given as a starting point.

Appendix: organisations working in the space of data ethics in the UK

Inclusion Criteria

Many organisations in the UK have published a one off report with some recommendations about the ethics of data in recent months and years. We include only those with a commitment to further work, through a strategy or sequence of activities, even if that strategy is not yet published. We also exclude those whose focus on data is restricted to high level uses of artificial intelligence (AI) or deterministic uses of algorithms. We also require that the organisations have a specifically ethical consideration of data use, i.e. what is right rather than simply using it. The general intention is to include bodies in this space, so the outward relations of bodies in some way is expected but not prescribed.

Information Commissioner's Office

The Information Commissioner's Office (ICO) upholds information rights in the public interest, promoting openness by public bodies and data privacy for individuals. Thy have recently expanded their team to include expertise on data use including machine learning.

Centre for Data Ethics and Innovation

A new independent government advisory body that will investigate and advise on how we govern the use of data and data-enabled technologies, including artificial intelligence. Examples in other countries have been fixed term so this is the first such body in the world and is expected to be put on a statutory basis in future.

National Statistician's Data Ethics Advisory Committee

NSDEC considers project and policy proposals, which make use of innovative and novel data, from the Office for National Statistics (ONS) and the Government Statistical Service (GSS) and advises the National Statistician on the ethical appropriateness of these. Now serves as ethical board for research applications to the Digital Economy Act (2017) and has developed a self-assessment tool for use by the ONS Data Science Campus. Also supported by a new research accreditation panel for data access approvals.

Government Digital Service

Produced the original Data Science Ethical Framework to guide the use of data in government applications, including various examples. Subsequently updated to a workbook and a Data Ethics Framework which will continue to iterate.

Confidentiality Advisory Group

CAG was established by the Health Research Authority to process applications for new uses of patient data held under common law understanding of confidentiality. Provides decisions in cases where returning to individuals for consent is not considered feasible, including Section 251 of the Health and Social Care Act (2012).

Secure Anonymised Information Linkage (SAIL)

SAIL is the Welsh data linkage project that provides access to linked government data for research and other purposes. Includes a public advisory panel which determines what is within the bounds of public expectations for the use of data.

METADAC
Medico-legal and ethico-social interdisciplinary data access committee for biological data associated with large UK longitudinal studies. Currently funded by research organisations to process research applications and develop public engagement and governance model. Developed principle of plain language summary to evidence research applications and need for broad interdisciplinarity in ethical approvals panels.

Digital Ethics Lab
Tackles the ethical challenges posed by digital innovation, helping design a better information society: open, pluralistic, tolerant, equitable, and just. Has a goal to identify the benefits and enhance the positive opportunities of digital innovation as a force for good, and avoid or mitigate its risks and shortcomings. Part of the Oxford Internet Institute.

Data Ethics Group
Part of the Alan Turing Institute, the UK national centre for research on Data Science and Artificial Intelligence. Made up of academics specialising in ethics, social science, law, policy-making, and big data and algorithms, the Data Ethics Group drives the Institute's research agenda in data ethics and works across the organisation to provide guidance on ethical best practice in data science.

Ada Lovelace Institute
An independent research and deliberative body of the Nuffield Foundation, with a mission to ensure data and AI work for people and society. Ada offers expert, independent commentary on the ethical and social implications of data, AI and related technologies, to inform the thinking of governments, industry, public bodies and civil society organisations in the UK and globally.

Wellcome Trust
Understanding Patient Data is the current focus of the Wellcome Trust data strategy, supporting better conversations about the uses of health information. Their aim is to explain how and why data can be used for care and research, what's allowed and what's not, and how personal information is kept safe. They work with patients, charities and healthcare professionals to champion responsible uses of data.

Digital Catapult Machine Intelligence Garage Ethics Committee
Small organisations such as start ups in the digital sector don't have the capacity to maintain their own ethics function. Therefore the Digital Catapult has established an independent ethics committee of its machine intelligence garage which is developing procedures for external approvals.

DataKind

A not-for-profit supporting third sector organisations to realise the value of their data through: ‘office hours’ consulting, data dives and larger pro bono consulting. Developed a set of ethical principles for data scientists who volunteer with them which they now plan to iterate.

Open Data Institute
The Open Data Institute works with companies and governments to build an open, trustworthy data ecosystem, where people can make better decisions using data and manage any harmful impacts. Produced a ‘Data Ethics Canvas’ which is iterating through experience. Now funded by DCMS to develop pilots of the fabled ‘data trust’ which can facilitate new models of data sharing.

DotEveryone
Doteveryone champions responsible technology for a fairer future and sees data within the context of its use in powering technology.

Datum Future
A not-for-profit think tank exploring the opportunities and challenges in the new world of data with a strong representation of business in contrast to government, charitable and academic initiatives above.