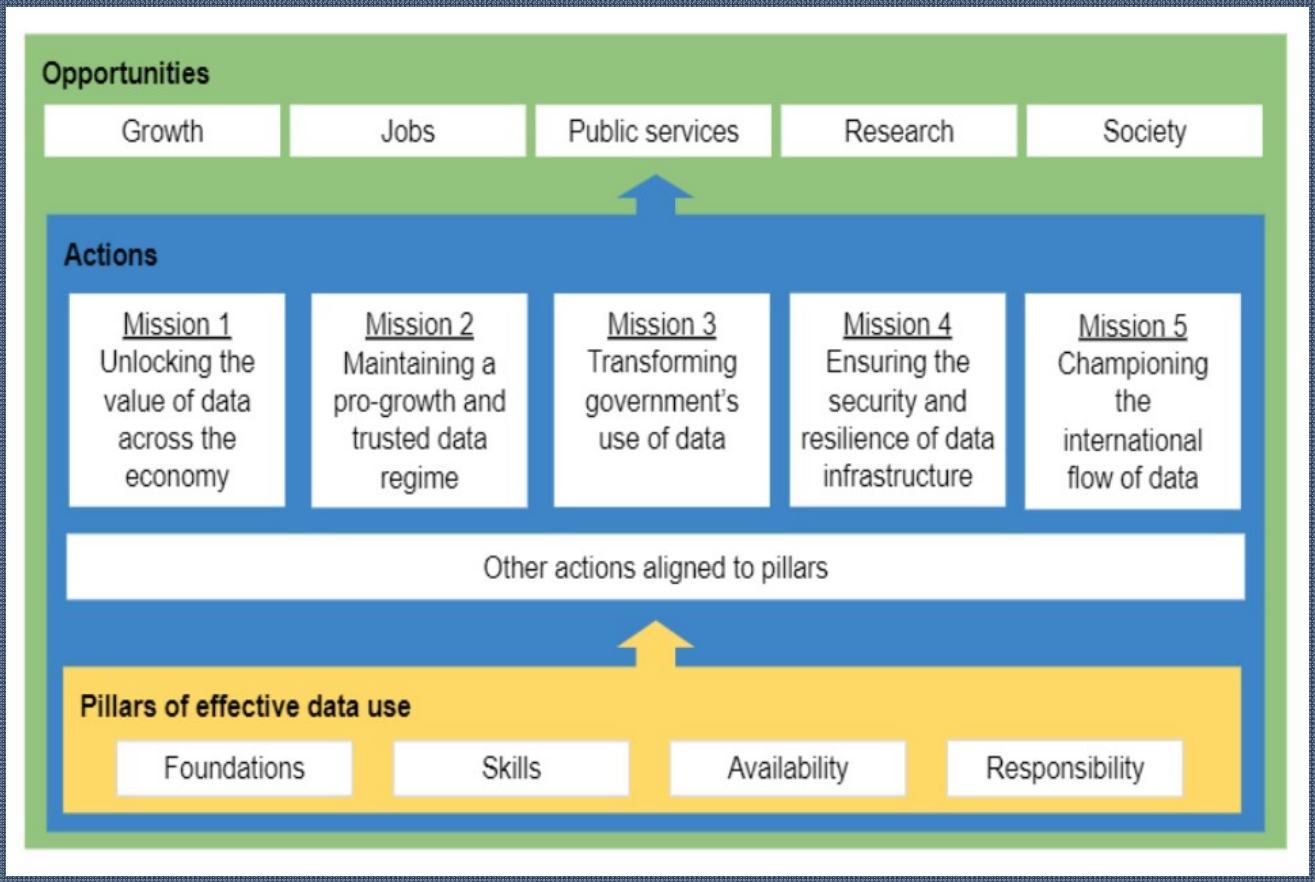

University of Edinburgh
UK Data Strategy consultation

THE UNIVERSITY of EDINBURGH

Q1. To what extent do you agree with the following statement: Taken as a whole, the missions and pillars of the National Data Strategy focus on the right priorities.

Please explain your answer here, including any areas you think the government should explore in further depth.

The University of Edinburgh considers the high-level priorities indicated by the missions and pillars of the UK data Strategy to be sound.

The University aspires to be one of the leading contributors in delivering the vision of the Data Strategy, through the large-scale Data-Driven Innovation (DDI) Programme. The aspiration of the 15-year DDI initiative, funded by UoE and both the UK and Scottish Governments as part of the Edinburgh & South East Scotland City Region Deal, is to position Edinburgh as the Data Capital of Europe. The five DDI Innovation Hubs (Bayes Centre, Edinburgh Futures Institute, Usher Institute, Roslin Institute, and National Robotarium with Heriot-Watt University) are externally facing innovation centres, supported by the data storage and processing capacity of the world-class Edinburgh International Data Facility.

The key opportunity is to ensure data is available for research & innovation in the UK across all domains. R&I has to be weaved in to all of the pillars and mission domains of the data strategy, including previous work such as the Smart Data Review and the Open Energy initiative.

Data access should be organised within secure, responsible, purpose-driven data environments, such as the Edinburgh International Data Facility (EIDF, part of the Data-Driven Innovation initiative) and the Global Open Finance Centre of Excellence (GOFCoE), recently set up at the University of Edinburgh, utilising EIDF. Data can provide enormous benefit and growth for the UK, but without a determined effort to make data available for innovation, much of that opportunity may go to waste.

The missions and the pillars of the UK Data Strategy align well with the five main themes and the vision of the Data-Driven Innovation (DDI) programme, as follows:

DDI Theme	Description	UK Data Strategy
TALENT	Creation and retention of a sustainable DDI talent pool, addressing current skills deficits and equipping more people with the skills needed to unlock the benefits of data for existing and new businesses.	P2: Skills
RESEARCH	Protecting, enhancing and expanding leading data science research to attract the best talent, and to continue to drive innovation and better connect the talent pool and industry to leading research.	P1: Data Foundations M2: Pro-growth, trusted data regime
ADOPTION	Increasing the use and application of DDI to accelerate productivity benefits, equip organisations (public, private and third sector) to harness data as a core asset, catalyse the market for new, better and less expensive products and services and enable the wider use	P4: Data responsibility M1: Value of Data Across Economy M3: Transforming

	of data to drive insights, benefits and inclusive growth.	government's use of fata
DATA	Maximising the value and usability of datasets and turning data into "assets" by improved data gathering, sharing, storage and analytical analysis and creating public trust in the data-driven economy.	P3: Data Availability M4: Secure and Resilient Data Infrastructure M5: Championing the international Flow of Data
ENTREPRENEURSHIP	Creating the institutions, networks and supporting environment to improve all modes of cross-sector partnership and collaboration (between private, public and third sectors and universities) to enable new and better DDI-based business models.	M5: Championing the international Flow of Data

"Doing Data Right" is the key aspiration of the Data-Driven Innovation Programme. We particularly welcome and commend the government for making ethical, sustainable and responsible data use one of the key pillars. The Strategy rightly identifies public trust in data as a critical vulnerability and concern for the future of data-driven innovation in the UK, and sees its link to responsible and ethical data practice.

A Good Strategy is an Actionable One

However, despite the sound priorities of ethical and sustainable data use, the document as a whole does not yet articulate a concrete, implementable strategy for addressing them, particularly concerning the responsible data pillar.

A successful strategy must identify not only goals, but also key challenges, tensions, conflicts and barriers to meeting them, and plans, resources and other means that the strategy will use to overcome these. The National Data Strategy identifies the right goals, but more work is needed to articulate the specific obstacles, and solutions and actions to address them.

Especially, the Responsible Data pillar of the strategy is noticeably muted or silent about the tensions between the acceleration of innovation/research and responsible use of data. In what sectors these are likely to arise, and how can such tensions be resolved in practice?

For example, the strategy misses the opportunity to address the public concerns about the government's use of data in response to COVID-19: debates over centralised vs. decentralised contact-tracing apps, discussions about the management of public health data, and questions about the public procurement of private technical expertise.

Of course, it would be a mistake to treat innovation and responsible data as inherently opposed goals. As the document makes clear, only responsible and ethical data use can earn the levels of public trust necessary to sustain innovation and adoption of new data-driven technologies. Nevertheless, it remains the case that the current market incentives for data-driven innovation are often misaligned with the incentives for responsible data use, and these misalignments need to be more openly and comprehensively grappled with in the National Data Strategy. The strategy should offer a concrete roadmap for the realignment of these incentives to enable the joint achievement of the pillars.

In order to do this, the pillars need to be addressed in a more holistic manner. The pillars are said to be 'highly interconnected'; but it is essential to show what the interconnections are, how they work, and their mutual dependencies and directions of influence. For example, the pillar of 'data skills' leaves out the broader context of education beyond the learning of 'skills'; i.e., the knowledge, understanding and wisdom that are

necessary to make the skills work to promote responsibility. This would suggest multidisciplinarity, lest the 'skills' pillar results only in highly trained data mechanics. Shouldn't we talk about, and develop, 'ethical skills' in data education? Among the questions for this would be: what kind of ethical curricula should be developed, for whom, based on what materials or models already in use, how should it be resourced and disseminated, by whom, and with what level of disciplinary granularity and focus?

Sec. 7.1.2 is about fairness, transparency, and trust, showing low levels of public trust in government data use but is short of saying how these can be increased, except for referring to the need for a 'sound ethical and legal framework' to engender trust. What would that framework look like? The document rests heavily on the CDEI's actual and potential role. However, there are unanswered questions about CDEI's future institutional position, independence, organisational relationship with parts of government that do public policy, resources and capacity to deliver.

Data Protection does not Guarantee Ethical use of Data

The current Strategy should be reformulated to avoid two dangerous equivalencies that are made or implied throughout the document. The first is between ethical or responsible data use and *data protection.* These are related and mutually reinforcing but quite distinct spheres of action, and the UK cannot achieve the goals of data ethics by merely attending to the technical and legal requirements of data protection. However, the Strategy frequently conflates them, as in Mission two, which states that the government is committed to "high data protection standards so that data processing is fair and does not result in discriminatory outcomes."

Two separate concerns fuse here in a worrying fashion. Data protection does *not* ensure fairness in data processing, modelling, or deployment, nor prevent discriminatory outcomes. Even improving data *quality* does not reliably assure fair, equitable, or non-discriminatory data/AI models and deployments. The Strategy needs to be informed by a ***more apparent distinction between data protection and data ethics, and a fuller account of the well-known causes of unjust, unfair, and discriminatory data practices and outcomes.***

Without Human input, Data is not Knowledge

A second dangerous and false equivalency is the statement that "Data is knowledge." (2.1) This view is dangerous as it equates data with obtained truth, when in fact, data are often false, corrupted, incomplete, biased, mislabelled, or missing vital context. An immense amount of painstaking and rigorous human work—data curation, cleaning, labelling, validation, interpretation, interpolation and testing—must be added to data in order to have any hope of making reliable knowledge from it. Responsible and trustworthy data scientists know this, and never equate data with simple truth or direct knowledge. It is essential for public trust that the government convey this humble reality as well.

Mature Data Culture Understands and Respects the Risks

Finally, the National Data Strategy should avoid attributing the challenges of data-driven innovation to the existence of public attitudes it reflexively frames as unreasonable. For example, the primary obstacle to a 'mature data culture' is described as an "overemphasis on the challenges and risks of misusing data." By equating concerns about the risks and harms caused by data misuse with "immaturity" this language undermines rather than facilitates public trust in the Strategy and data. Public trust requires the public to believe that the government regards data maturity as *co-extensive* with responsible concern for data's genuine risks. A mature data culture would also include, as part of data skills, a more comprehensive social distribution of the capabilities needed to identify and challenge misuses of data. This language has sent the opposite signal, namely that the government wants to *reduce* attention to such risks and the dangers they pose to the public.

This is reinforced by the Strategy's framing of "the effort to identify and assess the quality of data sets" and "access constraints for commercial users" in the health data domain as *barriers* (presumably to be removed), rather than *responsibilities* that are essential to the successful use of, and public trust in, new data

applications. That trust depends on the demonstrable *trustworthiness* of government and innovation, not on the circumvention to which the Strategy often resorts.

The Strategy repeatedly references 'risk aversion', 'data-hoarding' and other pejoratively described behaviours as obstacles to innovation. It does not confront the root causes of these attitudes and behaviours which may sometimes be unreasonable or unwarranted, but are very often rooted in legitimate concerns about data-driven harms that have *already* been permitted to undermine public welfare and safety, without robust accountability or recompense for those harms. Rather than implicitly labelling such attitudes and behaviours as obstructionism, the Strategy should adopt a more considered and nuanced approach to understanding the various drivers of these obstacles and their potential remedies.

As a further point, we advise the avoidance of language describing data as 'underexploited.' (4.1) Framing data as a resource to be 'exploited' amplifies public distrust. Data are, increasingly, parts of/reflections of people and their lives. The government's goal should not be to 'exploit' people's data but to care for and apply it responsibly for public benefit.

Q2. We are interested in examples of how data was or should have been used to deliver public benefits during the coronavirus (COVID-19) pandemic, beyond its use directly in health and social care. Please give any examples that you can, including what, if anything, central government could do to build or develop them further.

For question two, we are only looking for examples outside health and social care data. Health and social care data will be covered in the upcoming Data Strategy for Health and Social Care.

Global Open finance Centre of Excellence (GOFCoE): Covid-19 data sharing

As part of the broader DDI Initiative, the Global Open Finance Centre of Excellence (GOFCoE) has secured five-year funding from the UKRI Strength in Places Fund to become the global leader in Open Finance.

GOFCoE aims to rebalance financial services markets in favour of the consumers and firms using these services, undertaking data-driven research to deliver social and economic benefits. GOFCoE is a world first, providing leadership, coordination, research, and capability to develop the benefits of Open Finance and to help safely unlock the potential of customer data as a force to improve lives. GOFCoE is an industry, regulatory and academic collaboration, focusing on areas of the market that can only be really solved by joining forces.

Some examples include:

- Data Sandpit of pseudonymised financial data contributed from multiple market verticals to help banks, fund managers, insurers and Fintechs to more rapidly develop hypothesis, prototype algorithms, test business models and work with regulatory sandboxes to develop proof points.
- Global Economic Observatory - a longitudinal study of how humankind earns, spends and saves through the lens of both consumers and businesses. This observatory will draw a wide variety of private sector and public sector data sets to provide an unprecedented research and policy capability including looking at things like how people manage unfair credit or prepare for long life.
- Algorithmic Bias Test Laboratory - a new capability to assist financial services practitioners to reduce discrimination, providing assurance of compliance and ethical standards in their algorithmic distribution of products and services.

- Global Open Finance Technical Standards Working Group - creating a digital library of the output of national and international agencies developing API standards, coupled to a working group of those agencies seeking to develop harmonisation of standards and interoperability across markets.
- Global Anti Money Laundering Data Collaboration Group - providing a highly secure data collaboration environment to enable enhanced pattern recognition in financial data sets and the combating of organised crime and terrorism.

Covid-19 Data Analysis

In late March 2020, in response to the pandemic, the GOFCoE team initiated a Covid-19 economic analysis project, with Citizen and SME workstreams, to use private-sector financial data (Banks, Fintechs, Credit Reference Agencies, Cloud Accounting providers) to support public benefit research.

GOFCoE engaged a wide network of private finance sector data suppliers to provide de-identified data voluntarily with no payment in return. The team sought multiple sources of data to ensure a representative picture of the economy whilst applying data minimisation principles (without using "Open Banking" data), developing a seven-step legal and risk-based approach for data sharing. They used privacy-by-design principles to put in place Data Sharing Agreements with providers and Information Sharing Agreements with information recipients (UK Government departments, including BEIS, HMT, BoE FCA). The data suppliers were responsible for all de-identification. However, the GOFCoE team perform checks throughout the data lifecycle to mitigate re-identification risks, underpinned by privacy-risk protocols. To date, two private sector organisations have been supplying data: a bank for the citizen project and a cloud accounting provider for SMEs.

On the citizen project, financial data has been aggregated to provide summary measures and breakdowns over time including, for example, income broken down by salary, benefits, pension, investments, and interest, with similar expenditure metrics. There will be material insights for UK Government from this data. Currently, the SME supply of data is not sufficiently representative to provide material insights at this stage – the team continues to work towards more data supply arrangements.

The output is presented in the form of UK hexagonal maps at the postal district level, on several dimensions, e.g. gender and age. These dashboard insights allow the UK Government to "slice and dice" the data to develop insights towards targeted policies.

It is still early days in the GOFCoE start-up process, but the centre has already demonstrated reasonable progress in a short space of time, given a complete standing start in March.

Problems encountered

Initial engagement from companies was very positive, with company representatives saying there was a willingness to provide data voluntarily to support this project for the benefit of the UK, in a time of global crisis. In many cases, this initial enthusiasm unravelled through follow up conversations. Broadly the reasons were as follows:

- Data providers had their business priorities in response to the pandemic - for themselves as a business, their staff and their customers. Making data available requires expertise that most companies have in short supply.
- Some data providers were keen to engage directly with Government, eying both the opportunity to establish commercial relationships and build influence
- One data provider saw us as direct competition to their ambition despite our not-for-profit mission
- Follow up conversations with more willing participants frequently got bogged down in complex and nuanced areas, such as ambiguity regarding "consent" (as a legal basis versus permissions and

invisible processing), the definition of personal data (versus pseudonymised data), customer expectations, GDPR concerns as well as ethical concerns
- Dealing with the wrong person in the organisation, e.g., accessing sufficiently senior decision-makers or business development versus legal and data teams

In successfully concluding a Data Sharing Agreement with the Bank for the Citizen project, the University of Edinburgh accepted a sizeable commercial liability for the data. The Bank initially proposed it should be uncapped, then proposed it became £100m and eventually the contract liability ended up at £5m. This reduction was a direct result of being able to demonstrate our privacy by design approach with the right person in the Bank..

Reflections on the experience

1. **Mission led**. It is important to note, as discussed above, that the mechanisms to support data sharing are much broader than legal frameworks. The end-to-end journey to support the use of data for research & innovation is complicated, from culture to commercial motivations to legal and ethical impediments. This time, the binding force to overcome these difficulties was Covid-19. The next imperative is likely to be the Climate Emergency. However, there still is no sufficient top-down imperative from these two existential risks to unlock private data fast enough. Perhaps other measures are needed, including consideration of legislation that balances individual rights over data with the need to drive economic, public & social benefits. Each imperative must sit in the context of trusted public-private data sharing, forged in clear legal frameworks.

2. **Engagement.** As a start-up, the team realised they could have been even more focussed. Prioritising fewer, larger companies, with direct Government support to the ask, could have accelerated progress.

3. **Public Trust**. GOFCoE has always considered public trust as essential to delivering its mission. Collaboration across sectors to deliver better outcomes is crucial, including close relationships with the third sector. This reflects the GOFCoE approach to privacy and security, including active engagement with the Information Commissioner's Office (ICO) which has enabled the development of the proactive, robust risk-based approach to data and information sharing. Investing in the data governance and resulting assurance yielded "trust" benefits when dealing with the Bank.

4. **Friction** – Consistent feedback from UK Government departments was that direct data supply from companies was siloed and unrepresentative, making policy-making more difficult. Having a single source, like GOFCoE, would produce more representative insights that would remove much of this friction and accelerate delivery of critical economic insights at a critical time.

7. **Cost**. The cost and effort to engage in putting voluntary data-sharing agreements in place with the private sector is significant. The cost of commercial, legal, technical, IT infrastructure & security, governance controls to ensure GDPR compliance, analytics, disclosure control, insurance, et cetera, all adds up. GOFCoE is in a unique position with a mission-led objective to undertake such a project voluntarily. It relies on developing trust and goodwill between the University of Edinburgh and the private sector companies, as well as an environment of public trust in the use of data for research purposes - supported by legal and contractual frameworks. Suppose there is a public trust backlash (the Experian ICO case as a recent example). In that case, the private sector may become even more reluctant to provide this data voluntarily. There may be the need to centralise costs associated with these processes, perhaps through a small number of accredited (but independent) centres.

8. **Capabilities**. EPCC stores the Covid-19 project data securely, subject to extensive governance controls, with an ethos based on accountability, transparency and lawfulness, rooted in the principles of Privacy by Design (PbD). As a responsible custodian of data, GOFCoE builds its data management on robust legal, security and privacy risk processes, including ethical assessments for fair and proportionate data sharing.

GOFCoE views this approach as crucial to ensure public trust in how data is used and have some unique capabilities to enable this.

9. **Financial Liability** will remain an issue as we move forward. If GOFCoE has to accept additional liabilities, ultimately the ambition will be limited by the level of risk the UoE can accept. It would be useful to think mechanisms to manage such liabilities for Research or Public Benefit purposes, particularly in a crisis, and how to share such a burden across beneficiaries.

10. **Neutrality**. The availability of private sector data (in this case financial data) in a neutral, accredited location with the ability to combine it with other public sector data (e.g. health) in a **strictly controlled manner** for specific independent research & innovation purposes is essential for the UK to meet its goals including increased national productivity. Experience on the Covid-19 project highlights the thorny practical difficulties associated with unlocking this type of private-sector data. The GOFCoE team would welcome additional support from UK Government to sustain such economic and social research by making it much easier & faster to secure data (legal basis including consent) and by providing funding for this type of research.

11. **Culture around well-governed data sharing**. GDPR sets out how data can be shared, and this already includes research and public interest considerations. However, it is complex to implement and interpret data regulations. The law can be abstracted from implementation with significant gaps when it comes to upholding data rights or understanding where they are not absolute, or more broadly in the comprehension of data compliance and risk. Ethics and regulation can be used as a defence mechanism to justify not sharing data. In general, companies often find it easier to say no rather than yes to voluntary data requests for public benefit research particularly if the request is perceived to be onerous in terms of resources to fulfil the request.

12. Better outcomes. Essential reflection (or indeed hypothesis) is that Covid-19 will cost the UK ~£300bn. Had decent data sets been available earlier, the GOFCoE team could have contributed to a significant reduction in this number by enabling drive data-driven policy decisions. That certainly is a bold statement, but with the right data, a mission-led incentive, good processes and privacy practices, we could have been better prepared for this crisis as a nation. As it stands, we can still achieve some of this for the here and now (economic fallout will continue for the most vulnerable in society for years to come), but there is also merit in having a fine grain, longitudinal economic understanding which would serve us well to support research that could inform response in futures crises.

Q3. If applicable, please provide any comments about the potential impact of the proposals outlined in this consultation may have on individuals with a protected characteristic under the Equality Act 2010?

The document does not identify specific strategic concerns about these impacts beyond general references to ‘fairness’, ‘bias’, and discriminating against ‘different groups’. Nevertheless, equality impact assessments, and to an extent, data protection impact assessments, as well as algorithmic impact assessments, focus implicitly or explicitly on ‘protected characteristics’, and many ethical frameworks do so as well. An effective National Data Strategy must outline ***concrete steps to take to ensure the just, equitable, lawful and responsible treatment of protected persons and groups.***

Moreover, the Strategy should avoid the false dichotomy between protecting ‘individual rights and public benefit’, portrayed as competing aims to balance. Sometimes, there is a need to strike a balance between a particular individual right and a benefit to the broader public. The Strategy does not say how those balances will be struck, by whom, and with what transparency and accountability. ***In general, protection of individual rights is a public benefit in its own right.*** It is part of what makes a society cohesive, secure, and mutually beneficial. It is vital that our National Strategy not treat these as incompatible goods to be traded in a zero-sum game.

Immigrants and asylum seekers may be seriously affected by the way data and algorithms may be used by the Home Office and Border Force. The use of proxy variables to identify people with protected characteristics (such as race or religion), even in purportedly anonymised or aggregated datasets, needs scrutiny from ethical and legal standpoints. People with physical or mental disabilities may be excluded from participating in data-related occupations or face barriers in access to, and use of, data as citizens (e.g., using public services) unless their particular needs are taken into consideration in designing systems. The way data sources categorise population data and classify people may have unintended adverse effects when making decisions or assumptions. We know that historical biases and discriminatory judgments and practices are embedded in most large datasets used to classify and predict human behaviour. Such tools carry a substantial risk of 'baking in', normalising or further calcifying those discriminatory practices in the future. A National Data Strategy that fosters more ethical and responsible data cultures, as well as public trust in data-driven innovation, will need to have highly developed plans for preventing such harmful and unjust outcomes.

Q4. We welcome any comments about the potential impact of the proposals outlined in this consultation on the UK across all areas, and any steps the government should take to ensure that they take account of regional inequalities and support the whole of the UK?

Protected characteristics do not include socioeconomic class, and class inequalities might not reflect in geographic 'regional' inequalities as such. To the extent that government statistics are or are not collected to show class or geodemographic differences and inequalities concerning the proposals of this document, the collection policy and frameworks should be reviewed and reformed, and research (and research grants) targeted to bring these to light.

Data-driven tools have, in many places, led to problems of inaccurate and exclusionary profiling that enables what Virginia Eubanks (Automating Inequality, 2018) has identified as the systematic use of data-driven tools to (often unwittingly) punish the poor. The Strategy should commit to the study of such risks and the safeguards/oversight mechanisms needed to manage them.

Q5. Which sectors have the most to gain from better data availability? Please select all relevant options listed below, which are drawn from the Standardised Industry Classification (SIC) codes.

- Accommodation and Food Service Activities
- Administrative and Support Service Activities
- Agriculture, Forestry and Fishing
- Arts, Entertainment and Recreation
- Central/Local Government inc. Defence
- Charity or Non Profit
- Construction
- Education
- Electricity, Gas, Steam and Air Conditioning Supply
- Financial and Insurance Activities
- Human Health and Social Work Activities
- Information and Communication
- Manufacturing
- Mining and Quarrying
- Transportation and Storage
- Water Supply; Sewerage, Waste Management and Remediation Activities
- Wholesale and Retail Trade; Repair Of Motor Vehicles and Motorcycles

- Professional, Scientific and Technical Activities
- Real Estate Activities
- Other

Digital and data-led disruption is at the heart of the future growth. It will enable transformational change across the economy, in almost any sector. It has the power to break down existing barriers and create highly profitable, new business models faster than ever before. On the current list of the most valuable companies in the world, 7 of the top 10 are digital businesses.

Data & Digital disruption is a horizontal, cross-cutting force across industries, and new competitors often emerge from outside the existing group of companies (for example, the impact of Air B'n'B to the hotel industry). Therefore, a sectoral approach has limited relevance in picking up potential winners. All sectors will experience the disruptive impact of data innovation, and data-driven powerhouses will not limit their business to any single industry sector (for example, Google). "Execution is everything" - the ability and competence to drive change through data is a sector agnostic skillset. For example, listing "Information and Communication" as one vertical sector amongst others does not correctly recognise the disruptive impact of that domain across all of the other sectors.

Nevertheless, at this stage of the development, the UoE believes there is still merit in combining such a horizontal, ecosystem approach with a vertical sector strategy. The industry domains with the most potential for rapid transformation are the ones which are already in the process of digital transformation. They have previous experience in data analytics or have the potential to benefit from rapid automation. We have identified Information and Communication; Financial and Insurance; Human Health & Social Work; Agriculture; Transportation and Storage; Wholesale and Retail; Professional, Scientific and Technical activities; Education; and Manufacturing as such sectors. However, none of the listed industries is exempt from the disruption.

We have emphasised the importance of integrating Financial Services data with other data in our response to Q4 (> ***See Global Open finance Centre of Excellence (GOFCoE): Covid-19 data sharing*)**.

We want to identify here another essential domain since it requires strategic, a focused collaboration between Universities, the public sector and private industries across the whole of the UK, and internationally: Health & Social Care Data.

Health & Social Care Data

Linked national healthcare data benefit patients, the health service, and the wider population, as proven by success stories from Scotland, such as the DataLoch Covid-19 Data Collaborative of the UoE. However, current activities are built on specific projects with progress fragmented and frequently limited to single conditions or regions. There is a risk that realising value and public benefit from these resources will be squandered unless it is made more easily accessible. Obstacles that impede data access include legal, technical and procedural bottlenecks.

To create an aspirational 360-degree view of the citizen, we need to progress towards a wider integration of public sector and administrative data sets from different domains. We need realism concerning the efforts to achieve this (skills & capabilities, governance processes and technology & infrastructure). Often, large, long-term initiatives are prioritised over more focused or regional projects, failing to see their potential to scale up.

We should build on the success of regional initiatives and work towards a standard data model that will ensure the rich local and regional data generated from across primary, secondary and social care is available on all persons with an acute or chronic condition.

A critical design decision will be how to store this unconsented personal data in a research-ready form and how to access research cohorts. Our aim should be to move away from the per-project preparation of datasets.

Instead, we need a system where we can use prepared mappings to the national standard data model to rapidly link data based on queries expressed against the data model, instead of piecemeal across many local data sets.

Q7. To what extent do you agree with the following statement: The government has a role in supporting data foundations in the wider economy. Please explain your answer. If applicable, please indicate what you think the government's enhanced role should be.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

The University of Edinburgh strongly agrees with the importance of the Government in supporting data foundations in the wider economy. The role of the government is critical to ensure data is available for public good: protecting open access for publicly funded data sets, ensuring data sharing between different layers of the government, developing smart regulation, preventing the creation of monopolies, agreeing data sharigj agreement s between the public and private sectors, and levelling the playing field to enable the access of SMEs to the data-driven markets – to name just a few operations.

Q10. How can the UK's data protection framework remain fit for purpose in an increasingly digital and data driven age?

In section 7.1.2 we lay out the functions of the Centre for Data Ethics and Innovation (CDEI), set up in 2018 to advise the Government on the use of data-driven technologies and AI.

Concerning the growing calls from the public and other stakeholders for more ethical data innovation safeguards, it would be prudent to use the toe-holds afforded by the ethical elements in the General Data Protection Regulation, the Law Enforcement Directive—both incorporated in the Data Protection Act 2018—to help turn data protection more towards ethically based regulation, rather than exclusively legal rules-based regulation.

Data Protection Impact Assessment (DPIA) might move more towards ethical impact assessment (and algorithm impact assessment) by asking questions about transparency, accountability, fairness, contestability, and safety/nonharm in the routines and templates that are used for DPIA. There is already some spadework in the literature on this.

Q11. To what extent do you agree with the functions set out for the Centre for Data Ethics and Innovation (CDEI) - AI monitoring, partnership working and piloting and testing potential interventions in the tech landscape?

The question could be more precise in its intent. Monitoring and study of AI's complex and evolving ethical challenges is an important area for public investment, and the appointment of an independent central public body charged with carrying out that work, if sufficiently resourced and empowered, could be highly beneficial. Such an agency can learn and adapt a working model from the many effective national agencies in the UK and elsewhere tasked with the study and regulation of aviation safety and medical device safety.

Such agencies have for decades served both the public interest and industry innovation in ways that are sustainable and aligned with growing public trust and confidence in new technology. The tremendous human and economic cost of their occasional failures (as with the regulatory capture by Boeing of the FAA in the

United States, leading to the 737 Max disaster) is indicative of the importance of such agencies operating with integrity and independence.

AI is not co-extensive with data-driven innovation. It is a substantial and growing component, but it would be a mistake to limit the scope of the CDEI to AI applications unless it is clear to whom the responsibility for oversight and promotion of ethical and responsible use of non-AI data applications will fall. Indeed, in the Strategy, the terms 'algorithms'/'algorithmic' are often used as if they are interchangeable with AI, or as synonyms for data-driven innovation, excluding other significant portions of the data ecosystem. It would be helpful for the Strategy to define its intended scope of the term 'AI' and situate it within that broader ecosystem of data and algorithmic applications.

Finally, the resourcing of CDEI is unclear. Can it afford sufficient staff with the knowledge and skills necessary to oversee private sector AI development? How will it acquire the powers to access the commercially proprietary information needed to monitor and test AI applications effectively? It is also unclear whether the question pertains only to public-sector uses of AI. Would CDEI have the knowledge and expertise to look across the whole public-sector landscape, from health to transport to law enforcement to the armed services? Again, this is a significant resourcing challenge from the standpoint of domain expertise and a governance challenge concerning authority and basis of the proposed partnerships. The compatibility or tensions between partnering and monitoring need to be explored.

Q11a. How would a change to statutory status support the CDEI to deliver its remit?

The independence of the CDEI from DCMS or other government bodies is a critical component of enabling it to deliver its remit and facilitate public trust. However, it is unclear how the Strategy plans to address the significant challenges involved and noted above, from accountability and authority to resourcing and staffing expertise.

Q12. We have identified five broad areas of work as part of our mission for enabling better use of data across government:

- Quality, availability and access
- Standards and assurance
- Capability, leadership and culture
- Accountability and productivity
- Ethics and public trust

We want to hear your views on any actions you think will have the biggest impact for transforming government's use of data.

All of these areas are essential, so it seems unhelpful to single out one as having the 'biggest impact,' since none of these can be foregone if the strategy is to succeed. However, regarding 'ethics and public trust,' the government has an opportunity here to make a critical shift in tone and approach, consistent with the strategy's encouraging acknowledgement that the government "must be willing to open itself up to scrutiny."

This is a vital observation. Despite the stated intent to "ensure that UK values of openness, transparency and innovation are adopted worldwide," the uncomfortable truth is that in comparison with many of its peers, the UK government has not been regarded as particularly open or transparent, abroad or at home. How does the National Data Strategy intend to make the cultural and institutional pivot to greater government transparency and accountability that will be necessary to achieve the goals of its programme of work on ethics and public trust in data? The strategy refers vaguely to the development of a 'robust ethical framework,' but does not describe how to secure this robustness. How will trustworthiness be engineered and demonstrated?

The strategy describes the government's commitments to increasing transparency and public engagement in a promising manner - improving "publishing of data by which progress can be measured", promoting wider use

of the Data Ethics Framework in the public sector, and supporting "data scientists and data policymakers to build lasting capability for ethical data use." More details about the realisation of these commitments would be welcome. We encourage the UK government to broaden its reach to the growing national and international community of experts in AI and data ethics who can help advise on these essential components of the strategy.

In particular, we advise that the UK government should seek out ways to invite more open and honest critique of data-driven initiatives in the public sector, and welcome external input and more rigorous 'stress-testing' of algorithmic or data-driven solutions to what often turn out to be problems far more complicated than any technical intervention can neatly 'solve.'

As evident with any number of recent examples from the Ofqual algorithm to the NHS-X contact-tracing app, it is vital for public trust that the uncertainties, limitations, value trade-offs and potential failure points of a data-driven approach are openly discussed, anticipated, and transparently managed with external inputs from both citizens and independent experts, even when new data tools are developed to manage intractable problems or an unprecedented crisis. Humility, honesty, and care are the virtues the government should endeavour to display in its deployment of these technologies, and at all costs to avoid overconfidence, anchoring biases and dangerous groupthink in the understandable desire to see innovative technical solutions succeed in the UK.

The strategy also makes no mention of the known harms, failures, and misuse of data in the public sector that the UK can learn from, both domestically and internationally. For example, several public unemployment and disability systems in the United States have been found by the courts to be misusing data and algorithms in ways that unjustly penalised, denied or reduced benefits to vulnerable citizens, in ways that were demonstrably arbitrary or based on woefully inaccurate data or analyses. (See the MiDAS case in the state of Michigan and the KW vs. Armstrong case in the state of Idaho for examples). How will the strategy create safeguards to prevent comparable future scenarios in the UK?

We commend the strategy for the many promising commitments noted above. Still, concerning investing more knowledge and resources in ethical and responsible data use, an effective National Data Strategy needs to do more than articulate an ambition to embed all good things into data-driven innovation. The strategy asserts that we will have openness but also security. We will lower barriers and friction to unlock the 'free flow of data', but also improve standards of ethical and responsible data use. We will push for productivity, and we will ensure accountability. We will build relationships of trust and cooperation but also take the 'opportunity to set the UK apart and take an independent, individual approach'. These need to be more openly acknowledged by the strategy as challenging problems, because they involve wrestling with what are, in our present systems, often divergent or offsetting incentives and strategic aims.

The strategy would be more likely to earn the public and industry confidence if it did include a detailed roadmap to outline the course that the UK intends to chart through the many challenges ahead. There are undeniable and significant tensions between potentially conflicting values of the data innovation ecosystem, regionally and globally, which need to be addressed in the strategy. These challenges and tensions are not insurmountable, but neither can they be sidestepped.

Q13. The Data Standards Authority is working with a range of public sector and external organisations to create a pipeline of data standards and standard practices that should be adopted. We welcome your views on standards that should be prioritised, building on the standards which have already been recommended.

Metadata standards for basic findability should be the primary focus of work in this area. In terms of standard practices, following FAIR principles would be a solid starting point. There are a myriad of data standards in

specific sectors. However, it is very important to ensure the UK picks up the speed and aligns with the international market where Europe, the USA and Japan have been showing leadership over the past few years.

A promising best practice is the Global Open Finance Technical Standards Working Group which is creating a digital library of the output of national and international agencies developing API standards, coupled to a working group of those agencies seeking to develop harmonisation of standards and interoperability across markets.

Q14. What responsibilities and requirements should be placed on virtual or physical data infrastructure service providers to provide data security, continuity and resilience of service supply?

The choice of protection of data provided by a data service provider is an important one and often not considered carefully by users of specific data services. While there are obviously minimum data security requirements for all data – even open data which at minimum must be protected from harm – the level of protection required for data should be a key choice made when the data owner is deciding where to place that data. For example, in the scientific research domain it is commonplace for data to be stored on "working" or "scratch" storage which is not backed-up. This works when the data can either easily be recreated or exists as one or more "golden copies" elsewhere (which will be backed up with full disaster recovery).

The key responsibility required of data infrastructure service providers should therefore be to let their users know what data security, continuity and resilience of service are afforded for data that uses their storage. As more and more data moves into Cloud services this information is not always easy to find and as a result too often data is stored in the Cloud in the assumption that it will be safe (securely held and backed up with disaster recovery measures). Clearer information on the services associated with data storage would be very helpful.

To be clear, not all data needs to be backed up or provided with disaster recovery services. What should be improved is the communication of how data will be managed when it is stored in a particular service allowing users to take informed decisions. At the moment, comparisons between services can be very complex.

Q14a. How do clients assess the robustness of security protocols when choosing data infrastructure services? How do they ensure that providers are keeping up with those protocols during their contract?

In general this is provided through the service agreements which are entered into when data owners choose which data infrastructure services they will use to store their data. As a data service provider, the University of Edinburgh regularly enters into such agreements. These agreements specify how the data will be stored and secured including detail such as network paths and access methods. It is commonplace to agree to physical inspections of the infrastructure and accommodation and regular IT Health Checks with penetration testing from externally appointed providers. As the contract often include significant indemnities, it is in the interests of the service provider to ensure they keep their security protocols up to date and regularly test them – daily security tests are commonplace.

Q15. Demand for external data storage and processing services is growing. In order to maintain high standards of security and resilience for the infrastructure on which data use relies, what should be the respective roles of government, data service providers, their supply chain and their clients?

As use of external data storage and processing services grows, the role of Government should be to set minimum standards and deal with issue of market abuse. For example, the European Commission has been working with Cloud storage providers to ensure they cannot charge for the moment of data out of a provider at the end of a contract and that the data is kept for long enough after the end of the contract to allow it to be reasonably copied to a new provider. The use of standard accreditations such as ISO27001 and ISO9001 for service providers is very important. It helps to maintain a level playing field and gives confidence to consumers that a sufficient level of maturity has been reached by the provider. In terms of the IT supply chain, modern procurement processes within the public sector ensure the IT suppliers abide by best practice in terms of security, maintenance and support of systems. In terms of commercial providers of data services, a Kite Mark system would be very useful to indicate to consumers that basic standards have been met by the provider.

Q16. What are the most important risk factors in managing the security and resilience of the infrastructure on which data use relies? For example, the physical security of sites, the geographic location where data is stored, the diversity and actors in the market and supply chains, or other factors.

There are two risk factors of critical importance. Firstly, the long-term storage of electronic data is a complex task with many associated risks. IT equipment is inherently unreliable and considerable skill is needed to design data infrastructures such data is reliably held and, when problems do arise, a regularly tested back up and disaster recovery plan can be applied. Much of these skills are learnt rather than taught and the UK should ensure it has a better programme of skills and training to develop the next generation of systems management and "DevOps" staff. We should not be relying on the USA and other countries to provide all of our storage needs – for resiliency the UK must encourage investment in this area. Secondly, the cybersecurity threat has grown exponentially in recent years. The most worrying aspect of this is its linkage nowadays to state actors. This is by far the biggest risk to data security and dwarfs worries about physical security and geographic location which are largely solved problems.

Q17. Do you agree that the government should play a greater role in ensuring that data does not negatively contribute to carbon usage? Please explain your answer. If applicable, please indicate how the government can effectively ensure that data does not negatively contribute to carbon usage.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

The University of Edinburgh **agrees strongly** with the need for the government to play a greater role in ensuring that increasing use of data does not contribute negatively to carbon usage.

Data is a powerful and essential tool in designing and delivering an effective response to climate change, and in reducing our climate impact in line with national targets and planetary needs. Integrating climate and environmental data is critically important to benefit from the insights data can provide: an integrated and balanced approach to data would help to redefine success measures, as a society and as an economy. By modelling cause and effect from climate actions to societal impact we can demonstrate the wider benefits of transitioning to a net zero economy, in particular on wellbeing, environment and biodiversity.

Currently, the lack of integrated data for responsive impact assessment on mitigation and adaptation measures leads to a lack of granularity and attribution with long lead times in correlating impact. At the same time, data sources and data capability to inform climate action are increasing exponentially. There is a clear need to gather and merge data on climate impacts into a single dataset and use this to increase the effectiveness of action and decision-making.

An integrated and balanced approach to data would help to redefine success measures, as a society and as an economy. By modelling cause and effect from climate actions to societal impact we can demonstrate the wider benefits of transitioning to a net zero economy, in particular on wellbeing, environment and biodiversity.

However, it would be absurd to fight cimate with data innovation if that would lead to exponentially growing energy comsumption, with waste heat and negative impact to the climate. We must embrace circular economy models to mitigate this rchallenge. Integrating climate action across the data strategy actions will compensate for additional costs, with significant opportunities (expenditure, jobs, exports opportunities) from recycling, redesign of products, energy efficiency measures in buildings, the application of circular economy principles, forestry and so on.

Q18. How can the UK improve on current international transfer mechanisms, while ensuring that the personal data of UK citizens is appropriately safeguarded?

> We will seek EU 'data adequacy' to maintain the free flow of personal data from the EEA and we will pursue UK 'data adequacy' with global partners to promote the free flow of data to and from the UK and ensure it will be properly protected.

The University of Edinburgh considers Europe (European Union) to be the clear priority location for data adequacy arrangements. Reasons for this are twofold:

- First, the European union is by far the largest teading partner of the UK, and is likely to remain so, regardless of Brexit (almost all countries trade motre with thei neighbours than contrues and regions wjhich are further away). The most important export product of the UK to Europe is services, and data adequacy arrangements are critical to continue fluent delivery of services to and from Europe.
- Second, legislation and regulation around key issues sich as privacy, IPR and data security vary a lot around the world. In the UK, the public opinion, public sector principles and values, and concrete legal arrangements of such key issues are largely compatible with Europe already. New arrangements are far more easy to organise between compatible data environments than incompatible ones.

Besides Europe, we need to ensure alignment with other global innovators, especially USA and Japan.

Authors

AI & Data Ethics (AIDE) Board: Prof. Richard Kenway, prof. Shannon Vallor, Gemma Cassells, Dr. Sarah Chan, Alistar Fenemore, Dr. Ewa Luger, Prof. Lesley McCara, prof. Mark Parsons, Prof. Charles Raab, Prof. Michael Rovatsos, Prof. Burkhard Schafer, Prof. Robin Williams

Global Open Finance Centre of Excellence: Damien McGarrigle, Kevin Collins, Sorcha Lorimer

Edinburgh Parallel Computing Centre EPCC: Prof. Mark Parsons

Jarmo Eskelinen, Director, Data-driven Innovation initiative (editor)