Written Evidence Submitted by Palantir Technologies UK Ltd (DDA0063)

Introduction

Palantir Technologies UK, Ltd. ("Palantir") is pleased to be able to provide evidence to the Science & Technology Committee's inquiry on "The right to privacy: digital data". In this submission, we seek to offer our high-level perspectives on how effective technology can support government and industry to leverage data across organisational boundaries, while simultaneously enabling high standards of data governance.

About Palantir

Palantir is a software company that makes data integration, analysis, and decision-making platforms. We employ over 700 people in the United Kingdom, with London our largest office globally and the research and development headquarters for our Foundry platform. Our business model is to license our platforms as software-as-a-service products. Palantir does not control, sell, or otherwise monetise its customers' data.

Palantir was founded in 2003, with an initial focus on the United States' national security community. There, our founders' objective was to enable information to be used across technical silos and fragmented organisations, while simultaneously ensuring the security of highly sensitive information and adherence to complex legal, privacy, and civil liberties requirements.

Nineteen years later, and following billions of pounds of R&D investment, our software can be found in every field of industry and government. In the aviation industry, Foundry supports a major aerospace manufacturer, its suppliers, and dozens of competing airlines to leverage a common data foundation, known as Skywise, dramatically improving aircraft manufacturing and maintenance. At the United States National Institutes of Health, Foundry powers the National COVID Cohort Collaborative Data Enclave ("N3C"), one of the largest collections of COVID-19 health records in the world, with over 8 million records from 65 separate institutions (and counting). And in England, the NHS has used Foundry to coordinate aspects of the COVID-19 response across the healthcare system – including the vaccination roll-out – and enable public transparency through the GOV.UK Coronavirus Dashboard.

In developing information technology systems for cross-organisation data utilisation, data governance needs to be treated as a first-order priority

Organisations commonly cite data governance requirements as a primary obstacle to the publicly-valuable utilisation of sensitive data across organisations. In our experience, this is not the case. The UK's data protection regime sets out clear rules and principles for processing of personal data, including the sharing of data amongst organisations. In general, these support public trust, consumer confidence, technological innovation and wider human rights norms. Increasingly, they are internationally-recognised.

Rather, most barriers to publicly-valuable cross-organisation data sharing arise from shortcomings in the tools and procedures that organisations use to manage their data, and a failure to consider data governance as a first-order priority in the design and deployment of information technology systems.

Ensuring good data governance across multiple organisations is technically difficult

In most large organisations, whether in the private or public sector, data is spread across a range of legacy IT systems that have accumulated over decades of mergers, acquisitions, and organisational restructuring. This complexity makes it difficult to answer basic questions about personal information and other forms of regulated data, such as: what data do we have and how is it used? How is it combined with other data? Where is it flowing to (both organisationally and geographically)? For what purposes is it being used?

The potential for failure grows exponentially as these disparate, custom-built solutions are layered on top of each other and required to interact. "Seams" emerge between systems, requiring data to be transferred manually (for example, in a spreadsheet attached to an email, rather than through an automated application programming interface), compromising the security of data transfers, breaking audit trails, and causing data controllers and other responsible parties to lose sight of what their data is being used for. It becomes difficult, for example, to ensure that data past its deletion date is deleted everywhere it has appeared across a system.

This is particularly acute in large organisations with broad mandates, and compounds when they seek to collaborate and share data with other organisations, outside the original data controller's direct managerial control.

Good data governance requires a comprehensive approach, and should be treated as a first-order priority for information technology systems – not as an afterthought

In developing information technology systems, particularly for sharing data across organisations, organisations should view data governance both as an engineering challenge and as a first-order design priority. They should not, as is common, seek to "tack-on" data governance capabilities as a secondary after-thought to the particular operational or policy function that the technology is supporting. Further, they should avoid commissioning bespoke software solutions, for specific one-off operational and policy purposes, without fully appreciating the data governance challenges that a fragmented architecture – with different data sources supporting different tools and applications – entails.

Central to the development of every information system, and every programme for sharing data across organisations, should be a sophisticated and comprehensive data governance foundation with tooling that empowers organisations, and data protection officers in particular, to control their environments. For example:

- it should be easy to understand where data is located, why, and what protections have been applied;
- where systems interoperate, records should be automatically generated about which data was transferred, why, and with what protections applied; and
- these records should be easy to understand, and, where appropriate, published.

These same imperatives apply where organisations and information controllers share data with other organisations.

So long as organisations and data-sharing programmes fail to treat data governance as a central and sustained design objective, there will always be a trade-off between effective data governance and achieving operational outcomes. Data governance considerations should be equally weighted to the ability of systems to meet operational objectives. If this does not happen, each system that an organisation adds will further fragment their data landscape, rendering effective data governance even harder.

The NHS Digital's General Practice Data for Planning & Research ("GPDPR") is an example of a public sector programme where these recommendations are important. GPDPR critics have argued

that the initiative represents a loss of control for patients: that there will be a lack of transparency into what patient data would be used for, and that opt-outs will be a laborious process (including printing out and delivering forms).

One way to give citizens greater trust in and control over the use of their healthcare data, as it is shared across organisations for different purposes, would be to provide them with a means of giving and withdrawing consent for categories of data use (e.g. 'academic research', 'system planning', 'private sector drug discovery' etc.). This could be as simple as allowing citizens to indicate their preferences through toggles on an app or website. Equally, patients could have detailed transparency into the different uses that their data was fulfilling.

However, with the data infrastructure relied upon technically unable to support this, this standard of data governance is unviable. This is not a failing of regulation - this is an infrastructure limitation leading to unnecessary tradeoffs: *'*research or control and transparency*'* rather than *'*research and control and transparency'.

Case studies

We have deployed our software as the infrastructure for a range of large-scale, cross-government and cross-industry data platforms, each spanning thousands of data sources and in some cases hundreds of distinct organisations.

The U.S. National Institutes of Health National COVID Cohort Collaborative Data Enclave (N3C)

> *The N3C has demonstrated that a multisite collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multiorganizational clinical data for COVID-19 analytics. We expect this effort to save lives by enabling rapid collaboration among clinicians, researchers, and data scientists to identify treatments and specialized care and thereby reduce the immediate and long-term impacts of COVID-19.*
>
> – Journal of the American Medical Informatics Association

One example is our work with the United States National Institutes of Health ("NIH"), providing Foundry as the software platform for the National COVID Cohort Collaborative Data Enclave ("N3C").

The N3C Data Enclave is accelerating research on COVID-19 by giving medical researchers access to the largest shared patient-level data asset of COVID-19 clinical data in the United States – and one of the largest in the world. Established at pace in the early stages of the pandemic, N3C now covers more than 9.4 million patients, from over 70 United States medical centres (and counting). Researchers can leverage more than 10.6 billion rows of data for approved research projects related to COVID-19.

N3C was established with data governance as a first-order design priority. The platform uses role- and purpose-based access controls to ensure that researchers only have access to data approved as being relevant to their role and particular research project. Medical centres contributing data to N3C retain visibility and control over how the data they initially collected goes on to be accessed and used, and so remain able to fulfil their obligations as data controllers. This compares with systems where data governance was not a first-order priority, and where data access can typically only be granted to the whole of a data source (and not to specific subsets of data, on a case-by-case basis), often with little or only extremely cumbersome auditing. Additionally, the platform provides anonymisation capabilities that prevent researchers from identifying specific individuals, and ensures that researchers can only analyse data within the platform – it cannot be exported.

The vaccine roll-out in England

> *Foundry has been a key tool in the NHS's biggest ever exercise in operational data integration (Gould et al 2020). It provided, in the repeated words of a number of those interviewed, 'a single source of truth' for what was happening within the vaccine roll-out. From where supplies were, to who had been jabbed, and where uptake was low. Based on that information, decisions could be taken and indeed options could be modelled.*
>
> - King's Fund, *The Covid-19 vaccination programme: trials, tribulations and successes*, January 2022

The vaccine roll-out has presented an unprecedented public health and logistics challenge. NHS England and NHS Improvement ("NHSE/I") have had to coordinate the creation of new supply chains and vaccination centres, for the administration of a new vaccine, at massive scale and short notice. This coordination spans hundreds of organisations, ranging from NHS organisations, GP practices and pharmacies, to Ministry of Defence units and road haulage firms.

Palantir's Foundry platform provides these many organisations, and approximately 8,650 unique users (around 2,100 of them using vaccine-related tools), with a common data foundation and a consistent source of truth. This has enabled the following outcomes:

- **Robust data governance:** Like the NIH, NHSE/I treated data governance as a first-order priority when establishing its data infrastructure for the vaccine roll-out. Access to data can be tracked and precisely calibrated on a need-to-know basis. These controls are applied at source, at the data controller's direction, and propagate automatically as data is transformed and deployed across different organisations involved in the roll-out. For example, a general practitioner's practice can only view data relating to their Primary Care Network ("PCN"), and only for certain pre-approved purposes validated through Foundry. This comprehensive approach, in which the governance regime propagates across a common data foundation used for a variety of purposes by different organisations, avoids the risks associated with fragmented processing across different systems, where access controls and audit trails are broken as data is sent between systems, and the visibility required to trace data quality issues to their source is lost.
- **Data-informed decision-making:** Personnel across the roll-out – in different functions and organisations – make decisions based on the same, near-real-time data foundation. They avoid the latency that exists where data is integrated on a case-by-case basis and shared in a disconnected way (e.g. spreadsheets circulated by email), and have a common, transparent starting point for collaboration and decision-making. They avoid time-consuming debates as to "whose data is most accurate", and fraught performance conversations where, for example, personnel in NHSE/I and a given PCN start from different views of the status quo.
- **Ready-for-use data, enabling the rapid and low-cost deployment of new capabilities:** Foundry has allowed the NHS to overcome the complexity normally involved in integrating, managing and securing data, to provide data that is ready-for-use across the roll-out. By comparison with a fragmented data landscape, where each new tool requires time-consuming and often consultancy-dependent data wrangling, this has dramatically accelerated the pace at which the NHS can deploy new capabilities. For example, within just three days, NHSE/I was able to establish a tool for PCN sites to report on site-readiness, enabling England-wide readiness assessment and the ability to investigate sub-performant sites or concerning trends.