The Alan Turing Institute

technopolis group

Review of Digital Research Infrastructure Requirements for AI

Foreword

Data science and artificial intelligence (AI) are changing the world. Inventions and tools that, not so long ago, were only science fiction are now a fundamental part of our everyday lives. The UK is a world leader in AI, and our researchers from across sectors are working on new technologies that will not only help address major societal challenges but also power economic growth and deliver day-to-day societal benefits for the public good.

The Digital Research Infrastructure underpinning the UK's AI ecosystem, comprising compute and storage facilities, data, tools, techniques, and people, is essential for AI researchers and innovators. It is a crucial enabler of research activities that vary from accelerating Machine Learning algorithms and developing Digital Twins, to supporting training programs and collaborative ecosystems. This evidence-based review outlines the current and future Digital Research Infrastructure needs for AI in the UK, in order to help us grow our national AI capability and make sure that the UK retains its world-leading AI status.

The study has identified four key findings: the UK needs to scale-up investment in Digital Research Infrastructure for AI; the compute for AI needs to be easily accessible, configurable, adjustable, and promote collaboration; any investment in hardware and compute for AI needs to be matched by investment in training and support; and unified data management standards and sharing policies need to be developed. If these recommendations are implemented in full, they have the potential to help cement the UK's position as truly world-leading and build on our excellence in AI research and innovation to date.

The demand for Digital Research Infrastructure in the form of compute, data and skills capability is set to grow significantly, especially due to a growing range of AI domains and emerging interdisciplinary collaborations. The key challenge for the UK now is to make sure that it provides all the necessary tools and support for its researchers and innovators to help them fully unlock the power of AI.

Sir Adrian Smith
Institute Director and Chief Executive
The Alan Turing Institute

Authors

Charlotte Glass, Technopolis

Tomas Lazauskas, The Alan Turing Institute

With contributions from:

Neil Brown, Reda Nausedaite, Felix Dijkstal, Aaron Vinnik, Bruno Raabe, António Neto (Technopolis)

Martin O'Reilly, Jennifer Ding, Arielle Bennett (The Alan Turing Institute)

Martin Hamilton (MartinH.Net)

Acknowledgements

The authors would like to thank the Steering Group and Advisory Group for their guidance and additional input to the study and critical review of draft outputs. The authors would also like to thank all the stakeholders who gave of their time generously to input to the study through interviews, workshops, and correspondence. The information, evidence and support provided were vital to the successful finalisation of this report. Lastly, the views presented in the report are the views of the community and are not necessarily shared by UKRI or any individual member of the Steering Group or Advisory Group.

Contents

Executive summary

The National AI Strategy[1] set out a ten-year vision to make the UK a global AI superpower and acknowledged access to people, data and compute as key drivers of progress and strategic advantage in AI. Digital Research Infrastructure (DRI) plays an integral role in the wider compute for AI ecosystem. Ensuring the DRI ecosystem meets the current and future needs of the research and innovation community developing and using AI will be essential to meeting the ambitions set out in the National AI Strategy.

On behalf of UKRI, The Alan Turing Institute in conjunction with Technopolis has conducted a review to better understand the UK's current and future DRI needs for AI. This exercise focused on consulting with AI communities, AI researchers and researchers who use AI to solve problems, plus wider stakeholders to understand their needs across three main elements (compute, data access and people/skills), currently and in five to ten years' time. This summary report sets out the views of this collective community, as communicated during the review.

Key findings

The UK needs to scale-up and then continuously invest in DRI for AI if it seeks to become a global AI superpower

Demand for compute and data capability for AI research has grown significantly in recent years and is expected to continue to do so, including throughout a growing range of AI domains. To address the future needs in DRI for AI, a long-term coherent programme of activities and investments will be required to support a scaling up and scaling out of compute provision, increased consolidation of data, the operational running of both compute and data facilities, the co-design and evaluation of new technologies, and the necessary training to support uptake and sustainability. This investment in DRI for AI would support the National AI Strategy's goal for the UK to be a "global superpower in AI" that is well placed to "lead the world over the next decade as a genuine research and innovation powerhouse."

This review shows that investment is necessary because the demand for AI computing capacity is increasing and existing AI-capable HPC centres are running at their capacity levels. This results in an imbalance between supply and demand as researcher demand for AI focused DRI increases, particularly from non-traditional computational fields.

Further, the compute capacity available for use within the UK at the national and regional levels is much lower than that available in the comparator countries. At present the UK does not have a national compute (Tier 1) capability for researchers wishing to use AI tools and techniques, limiting use for larger workloads. By comparison, France, Germany, Japan, and the United States all already have national AI compute capability in place and are continuing to invest at scale in next generation facilities. Additionally, whilst the UK scores highly on "talent" metrics in exercises such as the Global AI Index[2], it has a much lower score for a number of key areas including infrastructure, operations and commercial exploitation of research outputs.

Compute capacity for AI needs to be increased while ensuring it is easily accessible, configurable, adjustable, and promote collaboration to enable major scientific advances

Researchers primarily obtain AI compute capacity from their own labs and institutional level provision coupled with commercial cloud services. As there is no set approach in place for measuring national compute capacity that also incorporates institutional compute provision and access to commercial cloud, measuring overall compute capacity available for AI is a significant challenge.

Participants in the review generally supported the continued demarcation of compute provision using a tiered approach. However, infrastructure providers noted that at present the compute provision itself is often working at maximum capacity and fractured in its coordination and delivery. It follows that work to address barriers to access alone would increase use of existing AI-capable facilities that are already working at maximum capacity. The vast majority of researchers expect their compute needs for AI research and innovation to more than double in five years' time and indicated that access to computing systems with Graphics Processing Unit (GPU) accelerators is a priority.

Therefore, there is a pressing need to increase the compute capacity available for AI at different levels in a coordinated way that facilitates equitable access across the research and innovation community. This strongly aligns with the UK government's "place" agenda – lower barriers to access could increase diversity of participating organisations, support researcher engagement from nascent AI areas, and spur innovation in both foundational and use-inspired AI research. This report identifies three areas where targeted intervention around compute hardware may be transformative:

- **Tier 1**: Incorporate AI-capable research infrastructure such GPU accelerator hardware and cloud access models into the UK's next Tier 1 national scale compute service
- **Tier 2**: Uplift existing AI-capable Tier 2 facilities through further rollout of GPU accelerators and adoption of cloud technologies coupled with support for operating costs and continuity of service
- **Tier 3**: Encourage uplift of institutional compute provision to enable access to AI-nascent communities / students, as well as proof of concept studies

A coordinated strategy and associated support to ensure adoption of common / standardised software and tools such as cloud type approaches (e.g. container-based virtualisation) is needed to lower the barriers to access, enable interoperability and ease of movement between systems. This strategy should be applied to all three levels of compute provision without delay, however it is recognised that Tier 3 developments are generally led by institutional requirements rather than directed by government or research councils.

Any investment in hardware/compute for AI needs to be matched by investment in training and support to maximise uptake, efficiency and generated scientific outputs

The number of staff that help enable access to the existing Tier 2 and Tier 3 compute provision and the skill sets they have both need to be increased. This could be addressed by increasing the core funding available to Tier 2 facilities to cover operational expenditure, and provision of a dedicated resource for institutional DRI capacity. DRI providers at both Tier 2 and Tier 3 levels struggle to recruit and retain staff with the necessary skill sets due to pay scales and clarity of career pathways.

There is also a need to provide stable and continued support to Research Technology Professional (RTP) career paths and competitive pay structures within DRI facilities and institutional teams. Given the fast-paced development of AI tools and techniques and the need for truly continuous and ongoing professional development, there is a need for RTP upskilling programmes and resources which include AI.

A broad set of engagement and training programmes will be needed across the DRI ecosystem for the breadth of potential users. This will include activities to raise awareness of capability, demonstrate the potential applications of AI across a range of different research fields, and support upskilling of users. Such activities and training programmes could helpfully be centrally coordinated across the DRI facilities to share resource and maintain consistency. Use of DRI for AI will also depend on a wide variety of training and support programmes for both academia and industry, beyond the scope of the DRI ecosystem to provide.

Unified data management standards and sharing policies are needed

The adoption of AI will depend on the development and implementation of standards and processes for collating, organising and sharing data for AI, in line with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles[3]. This is especially important for supporting data interoperability which is crucial for interdisciplinary research.

To support this, there is a need to encourage and incentivise the widespread adoption of data standards and best practices necessary for responsible use of AI in association

[1] UK Government (2021) National AI Strategy – https://www.gov.uk/government/publications/national-ai-strategy
[2] Tortoise Media (2020) Global AI Index – https://www.tortoisemedia.com/intelligence/global-ai/

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3. https://doi.org/10.1038/sdata.2016.18

with open data policies, for example, as requirements of UKRI grant awards. This also extends to existing public sector datasets, whose quality and suitability for AI could be greatly improved by building on expertise from organisations such as the Office for National Statistics (ONS).

Varied access and licensing models and data interoperability issues can make it challenging to combine data from multiple sources, and commercial datasets can be prohibitively expensive for researchers to license. There is, therefore, a need to explore mechanisms and instruments to support the collating of datasets from disparate sources and broker access to commercial datasets.

5-10 year outlook

If implemented in full, the recommendations put forward by the community, as identified in this review, could amount to an integrated and holistic programme of support for compute capacity, data access, and people and skills. This would likely have an important impact on the UK's ambitions to be world-leading in AI research and innovation over the next 5 to 10 years.

The key benefits envisaged include more straightforward and equitable access to significantly enhanced compute capability for AI research and innovation, supporting a wider diversity of research communities, organisations, and geographic locations. The enhanced AI capability would incorporate cloud native technology[4] where appropriate, and be complemented by a breadth of high-quality AI-ready open and public data sources. Improved arrangements would also be put into place for access to public sector data, restricted data and commercially licensed data.

In parallel, adoption of AI tools and techniques would be supported across research disciplines and in industrial R&D by developing and nurturing a highly skilled cadre of Research Technology Professionals and upskilling the wider research community. This would enable AI researchers to exploit DRI for AI to its fullest potential, through continued professional development, training opportunities and embedded support at an institutional level.

[4] Cloud Native Computing Foundation (2021) Official definition of cloud native technology – https://github.com/cncf/foundation/blob/main/charter.md

Summary of findings, requirements, and recommendations

Findings: Compute

Current state

- AI is already proving invaluable in addressing societal challenges in key areas such as Sustainable Development and the response to the COVID-19 pandemic
- Adoption of AI tools and techniques is rapidly accelerating and proliferating outside of core areas such as Computer Science
- Access to computing systems with GPUs is identified as the highest priority to meet the current and future needs of the AI community
- Demand for DRI for AI compute purposes will more than double over the next five years
- Researchers primarily obtain AI compute capacity from institutional resources coupled with commercial cloud services (over 50% of survey respondents)
- AI compute facilities at universities are often based in research groups and labs, rather than provided at an institutional (Tier 3) level

Barriers to adoption

- At present the UK does not have a national compute (Tier 1) capability for researchers wishing to use AI tools and techniques, limiting use for larger workloads
- AI-capable hardware at EPSRC's Tier 2 supercomputing centres is used, however these are already running at maximum capacity
- Researchers reported difficulty engaging with Tier 2 centres regarding AI projects due to convoluted access processes and unfamiliar technical environments
- Researchers are often constrained by capacity limitations of Tier 3 facilities and available budget for commercial cloud services
- Compute for AI research is not equally accessible across the research and innovation community

International comparators

- Comparator countries such as France, Germany, Japan, and the United States have national AI compute capability in place and are investing in next generation facilities
- The UK's petascale supercomputers cannot match the new generation of pre-exascale and exascale facilities launched by the United States, Japan, and EuroHPC JU
- The compute capacity available for use within the UK is significantly lower than that available in the comparator countries

Findings: Data access

Current state

- In spite of open data initiatives from UKRI and other public bodies, the potential of AI for research and innovation is not being fully realised due to issues around data access
- Researchers advise that broader adoption of AI will require significant effort around standards and processes for collating and organising data
- Data interoperability is becoming increasingly crucial to support use of AI techniques in an interdisciplinary research context
- The amount of data that researchers are working with is expected to increase tenfold over the next five years

Barriers to adoption

- Researchers advise that concerns about potential ethical, legal, and political complexities can have a significant effect on data sharing and re-use
- Data owners and users can struggle to prepare datasets for processing by AI tools due to lack of specialist expertise
- Varied access and licensing models and data interoperability issues can make it challenging to combine data from multiple sources
- Public sector data can be of varied quality and can be difficult to access
- Commercial datasets can be prohibitively expensive for researchers to license

Findings: People and skills

Current state

- Teams of Research Technology Professionals within universities are often relatively small and working to support a breadth of needs across the institution
- Central funding for DRI has often been capitalised, with little or no support for staffing to assist researchers in adopting and exploiting the infrastructure
- DRI providers can struggle to recruit and retain staff with the necessary skill sets due to issues around contract length, pay scales and progression opportunities
- Staffing may be constrained to particular projects, e.g. where a project has obtained funding for Research Software Engineering support

Barriers to adoption

- There are significant gaps in training and knowledge – 37% of survey respondents said they had poor or very poor skills in organising and structuring data and/or code
- Fast-paced development of AI tools and techniques highlight the need for truly continuous and ongoing professional development
- Training and documentation often do not reflect the differing needs of AI researchers (“tool builders”) and the wider research community (“tool users”)

Issues identified for further research

- Costs associated with researchers’ use of commercial cloud services for AI projects are unclear and may be significant when aggregated
- There is no set approach in place for measuring national compute capacity, let alone the capacity available for AI research
- Computer Science, Physics and Engineering were particularly strongly represented (two thirds of survey responses) - further work may be desirable to engage with the wider research community

Requirements and recommendations

Key: S = Short-term (up to one year), M = Medium-term (one-five years), L = Long-term (five-ten years)

Compute	**S**	**M**	**L**
Accelerate planning for national Tier 1 scale facility by feeding in review findings and requirements, ensuring that it is internationally competitive	●		
Undertake further work to measure the compute capacity available for AI in the UK, especially at the institutional level, to help ensure that investment is targeted appropriately	●		
Scale up and out the UK's existing compute capacity for AI, e.g. by expanding existing AI-capable facilities, establishing new ones, and/or purchasing	●	●	
Support the sustainability and continuity of existing UKRI supported AI capable systems, e.g. through grant extensions and/or recurrent funding for operational costs	●	●	●
Support increased coordination and collaboration among DRI providers, e.g. through initiatives such as DRI Retreats	●	●	●
Encourage development and uptake of tools improving the accessibility and consistency of DRI systems, e.g. notebooks and containers/virtualisation		●	●
Raise awareness of DRI for AI facilities, support availability and access models, e.g. through DRI directory and DRI ambassador network	●	●	●
Explore the potential of next generation AI systems through a technology foresight and horizon scanning initiative, e.g. building on ExCALIBUR testbed approach		●	●

Data access	**S**	**M**	**L**
Support development and adoption of data standards to ensure research data is AI-ready		●	●
Support for interdisciplinary AI research, e.g. by identifying core data storage and management requirements across disciplines	●	●	
Improve accessibility and quality of public datasets for AI, e.g. by supporting development of exemplar datasets and supporting tools/documentation	●	●	●
Explore mechanisms and instruments to support the collating of datasets from disparate sources and brokerage for access to commercial datasets	●	●	●
Explore the potential of co-locating compute and data for key large scale public datasets and Trusted Research Environments	●	●	●

People and skills	**S**	**M**	**L**
Explore potential funding models to support RTP career paths, competitive pay structures and job security within DRI facilities and institutional teams	●	●	●
Explore approaches to providing dedicated institutional DRI staff capacity		●	●
Provide opportunities for RTP staff upskilling in AI, e.g. through a programme of training and supporting resources, focusing on AI	●	●	●
Promote uptake of AI techniques in under-represented disciplines, e.g. through support for AI training courses with a domain/research field focus	●	●	●
Continue to support training of AI specialists to maximise their use of DRI, e.g. by building on initiatives such as Turing AI Fellowships and Centres for Doctoral Training in AI	●	●	●
Engage with industry and academia to raise awareness of the potential of AI, e.g. through upskilling and training programmes		●	●
Support for communities of practice and interdisciplinary collaboration, e.g. through demonstrators showing the potential of AI in nascent fields	●	●	
Support for DRI providers to engage with non-expert users, e.g. through training, guidance and resources	●	●	●

1. Introduction

This document presents the primary output of the Review of Digital Research Infrastructure Requirements for AI, conducted by The Alan Turing Institute in conjunction with Technopolis and on behalf of UKRI.

The Alan Turing Institute with support from Technopolis undertook an evidence-based review into the UK's digital research infrastructure (DRI) needs for artificial intelligence (AI). This exercise focused on consulting with AI communities across the R&D landscape (AI researchers, and researchers who use AI to solve problems, plus wider stakeholders) to understand DRI needs and requirements across three main elements (compute, data access and people/skills), currently and in five and ten years' time. The views of this community have been captured and set out in this report. Overall, this review is expected to inform decision making and provide evidence for future investments in support of data and AI technology development in the UK.

For this study, the definition of AI outlined in the National AI Strategy is used: "Machines that perform tasks normally performed by human intelligence, especially when the machines learn from data how to do those tasks"; while DRI is defined to include "Large scale compute facilities; Data storage facilities, repositories, stewardship and security; Software and shared code libraries; Mechanisms for access, such as networks and user authentication systems; and the People, users, and experts who develop and maintain these resources".[5]

This review sits alongside other recent and current studies on related aspects of the DRI ecosystem, including reviews on: the Large-scale computing: the case for greater UK coordination[6], Software, Skills and Computing Needs for UKRI's research community[7], and the Future of compute[8].

The study itself was undertaken between February and June 2022 and involved multiple strands of data collection and analysis of existing evidence:

Desk research and analysis of existing data and information relating to the UK's AI landscape and DRI landscape, as well as the AI adoption and research landscape in comparator countries (Canada, France, Germany, Japan and the United States).

A survey of the UK's AI research and innovation community, including the AI researchers who are developing algorithms, tools and software frameworks, as well as those applying AI to support their R&D activities. The survey was disseminated through over 40 channels (e.g. mailing lists and social media accounts for DRIs, The Alan Turing Institute, UKRI and its constituent Research Councils) and secured 287 usable responses.

Interviews with 70 stakeholders from across the UK's research and innovation community, including representatives from academia, government, digital research infrastructures and industry.

This report sets out the key findings from the Review, summarised according to the different elements within the system: Compute, Data Access and People & Skills, and a final chapter then brings together a set of overarching conclusions that encompass all aspects of the DRI ecosystem.

[5] UK Research & Innovation (2022) Digital Research Infrastructure homepage – https://www.ukri.org/what-we-offer/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/
[6] Government Office for Science (2021) Large-scale computing: the case for greater UK coordination – https://www.gov.uk/government/publications/large-scale-computing-the-case-for-greater-uk-coordination
[7] Software Sustainability Institute (2022)
Software and Skills for Large-Scale Computing: collecting evidence to develop a National Research Software Strategy – https://www.software.ac.uk/news/take-part-survey-ukri-communitys-software-and-computing-needs
[8] UK Government (2022) Future of compute review homepage – https://www.gov.uk/government/publications/future-of-compute-review

2. Compute

2.1 International Compute for AI

The understanding that AI leadership is necessary for economic and social benefit has accelerated AI development initiatives globally. As a result, most countries have developed and implemented strategic plans to support the development and deployment of AI. Computing power is an integral aspect of AI and access to state-of-the-art compute is essential to fueling these initiatives; however, it is often not considered holistically in policy making.

Large-scale computing is typically divided between public sector, academic and industry systems. In recent years, many countries have increased their scale of public investment into compute for AI to meet the growing needs for research and innovation.

The metrics for measuring compute capacity available for AI are, at this stage, a complex and underexplored area. As the OECD AI Policy Observatory has recognised, there is currently no widely used definition of what "AI compute capacity" is, nor a clear framework to help countries measure their relative access to compute capacity.[9]

To get a sense of the current AI capacity available in comparator countries, their largest and most recent investments in large scale compute facilities for AI are often instructive. In particular, the computing power from Graphics Processing Units (GPUs), which have specialised processing units with enhanced mathematical computation capability, are advantageous for AI workloads.

In the **United States**, Oak Ridge National Laboratory's Frontier supercomputer boasts 1.1 exaflops[10] of performance. The system is the first to achieve exascale, currently ranks first on the TOP500[11] list and is more powerful than the following seven TOP500 systems combined. The Aurora exascale supercomputer, to become operational later in 2022 at the Argonne National Laboratory, will provide ~2 exaflops and is estimated as a more than US$500m investment.

In **Japan**, the National Institute of Advanced Industrial Science and Technology (AIST) runs the AI Bridging Cloud Infrastructure (ABCI), which is the world's first large-scale Open AI Computing Infrastructure. ABCI was upgraded in 2021 and now provides a peak performance of 226 petaflops of power. RIKEN in Japan also hosts the Fugaku supercomputer which provides 442 petaflops performance and has only been recently superseded by Frontier.

In **Canada**, the national compute infrastructure is facilitated the Digital Research Alliance of Canada, which coordinates access to the national HPC capacity via the national advanced research computing (ARC) platform. The ARC provides researchers with access to its five major supercomputers each offering between two and six petaflops, operated by regional partners across the country. In addition, one of Canada's four national AI institutes, the Vector Institute, operates its own AI computing infrastructure, which provides 12.5 petaflops performance and is open to applications from AI researchers throughout the year. As of August 2022, Canada has no plans to invest in a national exascale machine.

The European High Performance Computing Joint Undertaking (**EuroHPC JU**), a joint initiative between the EU, other European countries, and private partners to develop a World Class Supercomputing Ecosystem in Europe, has announced JUPITER, the first European exascale supercomputer to be installed in 2023 in Germany and four new mid-range (petascale and pre-exascale) supercomputing hosting sites DAEDALUS, LEVENTE, CASPIr, and EHPCPL in Greece, Hungary, Ireland, and Poland. EuroHPC JU also just inaugurated LUMI, a pre-exascale 151 petaflops system located in Finland (budget of over EUR 144 million), currently ranked as the third fastest and third greenest supercomputer in the world, which is more powerful than the other four fully operational EuroHPC JU supercomputers combined (Vega in Slovenia, MeluXina in Luxembourg, Discoverer in Bulgaria, and Karolina in the Czech Republic).

[9] OECD AI Policy Observatory (2022) Measuring compute capacity: a critical step to capturing AI's full economic potential – https://oecd.ai/en/wonk/ai-compute-capacity
[10] Floating point operations per second ("flops") is a generally accepted measure of compute performance. Today's most capable supercomputers run at speeds of over an exaflop, or one quintillion (1018) floating-point operations per second
[11] TOP500 website – https://www.top500.org/lists/top500/2022/06/

Three further supercomputers are also shortly to be launched: LEONARDO in Italy, MareNostrum5 in Spain, and Deucalion in Portugal.

Besides participating in the EuroHPC JU, EU countries also continue to develop their own supercomputers as well, for example:

- In **Germany**, the JEWELS operated by Jülich Supercomputing Centre at Forschungszentrum Jülich as a European and national supercomputing resource for the Gauss Centre for Supercomputing. Capable of 70 petaflops.
- In **France**, the Jean Zay supercomputer will be upgraded to double its peak performance to 28.3 petaflops. Another new supercomputer, Adastra, will be deployed in 2022 with 70 petaflops of performance.

Finding: Comparator countries such as France, Germany, Japan, and the United States have national AI compute capability in place and are investing in next generation facilities

It is important to note that **China** has also been making significant progress in developing its own supercomputing landscape in recent years. According to the latest (June 2022) TOP500 list, it currently has the highest number of supercomputers in the world (173), which translates to 530 petaflops and 12 percent share of the total TOP500 list's aggregated performance. However, information regarding the details of supercomputers in China, as well as their usage, is somewhat limited and is therefore treated as beyond the scope of this review.

Globally, multiple public and governmental initiatives are also exploring how to support the adoption of cloud native technologies for research, such as the National AI Research Resource in the US, the Japanese GakuNin Cloud Adoption Support Service, the China Science and Technology Cloud (CSTCloud) and the European Open Science Cloud (EOSC).

These new systems and services will allow for new technologies and research via the power of computing and simulation methods as well as data analytics and AI. The extreme scale and performance levels allow for research that has not been achievable before as well as making the computing power more accessible to researchers, industry, and government organisations.

2.2 UK Compute for AI

The compute capacity available for use within the UK for AI within regional or national facilities is lower than that available in other countries.

As of June 2022, the UK has 11 computer systems in the TOP500 list used by academic and research segments, amounting to only 1.2% of overall TOP500 performance. These systems have traditionally been designed and used for modelling and simulation, however, more recently there has been an emerging interest and need to use such facilities for large-scale AI research.

HPC infrastructure in the UK consists of three tiers of resources, from the largest capability machines operating as the national service (Tier 1[12]) to regional / specialist hubs (Tier 2[13]) and local / institutional systems (Tier 3). The facilities naturally prioritise academic users, but industrial collaboration and use is encouraged. They therefore provide valuable compute resource and diverse computing architectures supported by local expertise.

The provision of compute for AI in the UK is currently primarily located within the Tier 2 and Tier 3 levels however the funding and delivery of these infrastructures are currently not coordinated. At the Tier 1 level, the UK's national supercomputer ARCHER2 does not currently host the hardware, such as GPUs, typically required for large scale AI workloads.

The Hartree Centre has been awarded £20M from the Department of Business Enterprise and Industrial Strategy (BEIS) for an AI-capable machine as part of the £210M Hartree National Centre for Digital Innovation[14] collaboration with IBM. However, this machine will not be fully operational until late 2024/mid 2025 and is industry-focussed – researcher access will be on a full economic cost recovery basis due to the funding model adopted.

Finding: At present the UK does not have a national compute (Tier 1) capability for researchers wishing to use AI tools and techniques, limiting use for larger workloads

The network of Tier 2 facilities incorporates systems that support AI to varying extents and with varying levels of uptake. They currently host around 26 petaflops of total peak performance, in addition to the 9 petaflops provided by the DiRAC's Tursa system. However, researchers' access to these facilities is affected by a range of factors, outlined later. These systems are considered technology-wise on a par with EuroHPC's petascale supercomputers, though still far below the performance provided by the exascale or even the new petascale systems. It is also important to note that **the current grant funding for all the Tier 2 systems is due to end between before late 2024**.

Finding: The UK's petascale supercomputers cannot match the new generation of pre-exascale and exascale facilities launched by the United States, Japan, and EuroHPC JU

Finding: The compute capacity available for use within the UK is significantly lower than that available in the comparator countries

The Tier 3 (institutional / university level) systems play an important role within the UK's compute infrastructure landscape, however, there is currently no centralised map or database of university level clusters. Within the scope of this Review, it has not been possible to fully map the number and capacity of smaller AI systems or commercial cloud provisions held by universities and institutions.

It is important to note that many of the UK's internationally recognised AI researchers also have industrial partners, such as Google (DeepMind), Facebook (Meta), Microsoft, IBM, who provide additional resources (both cash and in-kind) to fulfil computational requirements by providing access to their public commercial cloud or proprietary compute facilities. This is also unmeasured capacity whose importance is difficult to assess.

As a result of these industrial collaborations and uneven investments in Tier 3 and commercial cloud compute, it is evident that access to compute for AI research is not equally accessible across the research and innovation community.

Finding: Compute for AI research is not equally accessible across the research and innovation community

The UK also has several partnerships that facilitate access to European large-scale computing systems such as the Partnership for Advanced Computing in Europe (PRACE), ELIXIR, and the European Centre for Mid-range Weather Forecasting (ECMWF). However, the UK is not a member of EuroHPC JU and therefore will not have access to their pre-exascale and exascale systems to come online in the coming years.

UK researchers can access supercomputing facilities at Argonne and Oak Ridge National Laboratories in the United States through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) programme. However, such access is a competitive process, subject to the availability of compute recourses, and is not a substitute for a national compute facility.

In May 2022, UKRI announced the intention to prepare two strategic business cases, one focused on an exascale system targeting deployment by 2025 and one focused on investment in large-scale accelerator-based compute capability for the UK over the next few years.[15] This review will feed into UKRI and government preparatory work, including the Future of compute review.

2.3 Current use

This section presents a summary of **where academics currently go for their AI compute needs and current barriers relating to access to compute**.

There are many examples of world leading AI research which harness the power of the current UK's DRI at a large scale and could be further enhanced by making more powerful and more abundant computational resources available to them. This includes, for example:

- **Language modelling**: The University of Edinburgh research teams lead by Prof Mirella Lapata and Dr Kenneth Heafield focus on developing AI systems capable of advanced reasoning and able to draw conclusions from large and varied sets of data, and large language models for fast and high-quality machine translation, respectively.
- **Neurology**: The research group led Prof Parashkev Nachev at UCL works on AI

[12] DiRAC and ARCHER2
[13] Baskerville, Cirrus, CSD3, Isambard GW4, JADE2, Kelvin-2, MMM Hub, Nice (Bede), Sulis
[14] Hartree Centre (2022) Hartree National Centre for Digital Innovation website – https://www.hartree.stfc.ac.uk/Pages/Hartree-National-Centre-for-Digital-Innovation-(HNCDI).aspx

[15] UK Research & Innovation (2022) UKRI position on next phase of large scale compute investments – https://www.ukri.org/what-we-offer/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/ukri-position-on-next-phase-of-large-scale-compute-investments/

models to generate synthetic brain images by learning from tens of thousands of MRI brain scans taken of patients of various ages and with a variety of diseases.

- **Computer vision**: Project Odysseus at The Alan Turing Institute aimed at understanding London "busyness" during lockdown by collecting and analysing live data from JamCam cameras and traffic intersection monitors.
- **Bayesian Deep Learning**: development of tools to quantify uncertainty in AI and applied in various areas from automotive (autonomous driving, control, and computer vision) to medicine, and pioneered by research teams led by Turing AI Fellows Prof Yarin Gal and Prof Chris Holmes (University of Oxford), respectively.

Across the various user communities, compute needs vary widely, reflecting a diversity of workloads and use cases. Compute needs are also often determined by the data that researchers are working with and where / how it is stored, as well as ownership, privacy, and security needs.

All AI researchers and practitioners use a combination of different compute resources for their AI related work. However, **researchers in academia are mainly using their own compute resource or that available at their institutions for AI research**. Just over half of AI researchers use either their research group's, lab's or institutional compute as their primary resource (with 27% indicating that these are both their primary and secondary sources of compute). This reflects the fact that **the majority of AI development or application work, particularly early-stage development, is** currently **conducted on smaller scale local systems**. Interviewees also noted that AI researchers are often more inclined to use their group's or lab's funds to buy their own hardware for development work, so as to have timely access to equipment that they are able to configure to their needs, without needing to go through application processes required by compute hosting facilities.

Finding: Researchers primarily obtain AI compute capacity from institutional resources coupled with commercial cloud services (over 50% of survey respondents)

Interviewees also indicated that systems held within research groups are often tied to research funding and supported by PhD students. When such research funding comes to an end, the knowledge and capability may be lost. This loss can also create challenges for institutional level DRI support services and planning, for example around energy costs or storage, and the need for additional maintenance or operations support.

Finding: AI compute facilities at universities are often based in research groups and labs, rather than provided at an institutional (Tier 3) level

Around **half of respondents were currently using commercial cloud** (e.g. Amazon Web Services, Microsoft Azure, Google Cloud Platform) for their AI-related work, but **only 8% indicated that this was their primary source of compute**. Despite this, many interviewees predicted that the use of cloud will continue to increase in future, as it addressed researchers' needs for flexible, convenient access to compute, without lengthy proposal processes. Interviewees also indicated that cloud was particularly useful for prototyping and demonstrations, meeting spikes in compute demand (i.e. cloud bursting), or to meet specific hardware or software requirements. Although there are services in place to facilitate access to cloud (e.g. the OCRE cloud framework[16], coordinated by Jisc), greater reliance on public cloud provision could create additional challenges for researchers around data security, path-dependency and increasing costs.

Finding: Costs associated with researchers' use of commercial cloud services for AI projects are unclear and may be significant when aggregated

The extent to which AI researchers are applying for time on the existing Tier 2 systems varies significantly between facilities, depending on the systems and hardware available. Whilst under half of survey respondents were using Tier 2 services, only 21% stated that these were of primary importance to them. Interviewees highlighted that moving between systems at different tiers was a challenge, and 50% of survey respondents stated that compute provision did not align with their requirements. Interviewees showed limited awareness of what is available for AI research within the existing Tier 2 facilities and how this might meet their needs. Interviewees also reported that they had experienced challenges with portability / user interfaces and found access arrangements cumbersome. It is for these same reasons that AI researchers may choose to buy local compute or use cloud services more often or limit the scope of their work.

Additionally, most of the **Tier 2 facilities consulted for this study also reported working at maximum capacity**. Managers of these facilities indicated that, due to the prevalence of mixed workflows and cultural differences between communities, estimating the proportion of users of compute capacity working with AI was challenging. Broad estimates of uptake of these machines for AI ranged between 20% and 90% of users. Most Tier 2 facilities consulted do not currently have the capacity to meet the full breadth of current demand, let alone an increasing future demand for AI and conventional HPC capacity.

Finding: AI-capable hardware at EPSRC's Tier 2 supercomputing centres is well used, however these are already running at maximum capacity

Finding: Researchers reported difficulty engaging with Tier 2 centres regarding AI projects due to convoluted access processes and unfamiliar technical environments

Finding: Researchers are often constrained by capacity limitations of Tier 3 facilities and available budget for commercial cloud services

Thus, whilst the UK continues to punch above its weight in AI research according to metrics analysed in exercises such as the Global AI Index, it could be surmised that this is **despite** the mismatch between researcher requirements and service provision at all (Tier 1, Tier 2 and Tier 3) levels.

2.4 Future needs

The current tiered DRI system (i.e. tiered system of provision) works well to meet the diversity of needs across the research and innovation landscape and should be retained. In future, maintaining this tiered level of provision will help to provide access to a range of architectures for different user communities. This will include access to compute at a local level through to institutional, regional, national and international levels, where AI is one part of the systems available.

It has not been possible within the scope of this study to fully map the number and capacity of smaller AI systems held by universities and institutions. As these smaller scale systems are likely to constitute a significant portion of the overall compute available, this information gap limits the completeness of this initial exercise. The OECD AI Expert Group on AI Compute and Climate[17] is already working to identify the most appropriate approaches for measuring national compute capacity with the aim of creating a basic framework for understanding, measuring and benchmarking domestic AI computing capacity by country and region. UKRI should continue to support the development of this guidance and build on prior National e-Infrastructure Survey[18] work to ensure the UK DRI landscape is fully enumerated and understood.

Finding: There is no set approach in place for measuring or benchmarking national compute capacity, let alone the capacity available for AI research

Requirement: Undertake further work to measure the compute capacity available for AI in the UK, especially at the institutional level, to help ensure that investment is targeted appropriately

Most survey respondents indicated that their need for compute for AI for research and innovation would increase significantly (more than double their current usage levels) in five years' time, while 68% of survey respondents indicated that computing systems with GPU accelerators would be a high priority to meet their current and future needs. 61% of survey respondents also had special requirements for high I/O (input/output) throughput and 42% had requirements for low I/O latency.

For those that could provide estimates of future needs, their responses varied:

- **CPU Cores**: 35% of respondents estimated that they would need between 11 and 100 CPU cores for their typical workflow, while 18% estimated they would need between 100 and 5,000 and only 8% estimated they would need more than 5,000. In terms of their largest workflow, 22% of respondents estimated that they would need 11 to 100 CPU cores, and 23% estimated that they would need 101 to 1,000 CPU cores. Only 16% indicated that they would need more

[16] Jisc (2022) OCRE cloud framework – https://www.jisc.ac.uk/ocre-cloud-framework

[17] OECD Network of Experts on AI (2022) Compute & Climate – https://oecd.ai/en/network-of-experts/working-group/1136

[18] HPC Special Interest Group (2022) HPC-SIG publications – https://hpc-sig.org.uk/index.php/publications/

than 10,000 CPU cores for their largest workflow.

- **GPUs**: 29% of respondents estimated that they would need 5-16 GPUs for a typical workflow in five years' time, while 15% indicated that they would need more than 65 GPUs. In terms of their largest workflows, 21% estimated that they would need between 65 and 512 GPUs and 17% estimated that they would need more than 513.
- **Memory**: Estimated requirements for memory per compute node were widely distributed between 128 and 2,048 GB, with 15% indicating a need for more than 2,049GB for their largest workflow. For their typical workflows, 40% of respondents estimated they would need less than 128 GB.

These estimations should be taken with caution. Researcher predictions of their future use of compute can be fraught as their usage of compute will be tied to numerous factors including availability of compute, data, and people, research funding grants or working contracts, which play a significant role in shaping their research directions. As such, it is not surprising that survey respondents and interviewees often found it challenging to predict specific future compute needs. When asked to estimate their compute requirements in five years' time (in terms of CPU cores, GPU, or memory per compute node, for either their typical or largest workflows), **around a quarter of respondents did not know what their future requirements would be**.

By extension, making predictions of compute needs for ten years' time is significantly more challenging. It is not possible to provide a definitive quantification of the compute capacity needed for AI across the DRI system, however, these trends demonstrate that the need for compute resource will continue to grow at a significant scale.

As noted above, the compute currently available for AI in the UK is limited and running at maximum capacity. It is also important to emphasise not only the need for increased compute, but also that any increase in compute capacity is equally accessible and available to researchers across the UK.

As noted above, the use of cloud is common across the research community and likely to increase. Therefore, there is also a continued need for initiatives like the OCRE framework managed by Jisc, which helps researchers with procuring commercial cloud services. Whilst it is often more cost-effective to own infrastructure when computing demand is almost continuous, as detailed above, commercial cloud platforms do provide resources and capabilities of use for the research community, especially for heterogeneous workloads and cloud bursting.

Finding: Demand for DRI for AI compute purposes will more than double over the next five years

Finding: Access to computing systems with GPUs is identified as the highest priority to meet the current and future needs of the AI community

Requirement: Scale up and out the UK's existing compute capacity for AI, e.g. by expanding existing AI-capable facilities, establishing new ones, and/or purchasing commercial cloud AI capacity

Notably, although the existing Tier 2 facilities and institutional infrastructure may provide compute for AI, the extent to which they receive funds for operational expenditure varies. As the needs of AI researchers evolve, the need to support continual development work to maintain and upgrade AI equipment will increase. To make best use of the existing platforms and facilities, funding should be made available to cover operational expenditure of running Tier 2 facilities. These operational costs include both the hardware running costs and the costs of providing operational support to researchers and innovators to apply AI tools. This would also support broadening out the compute provision and skills base for AI within these facilities and across the UK. This support would also ensure the existing capability within Tier 2 facilities is not lost.

Institutional level facilities should also increase their levels of operational expenditure, although it is acknowledged that the running and delivery of these systems is primarily the responsibility of the institution. However, there is a need for a mechanism to support greater continuity of support for the delivery and maintenance of compute provision at the research group level, which should be considered by both UKRI and institutions. This may include maintaining stronger records at the research group level, or institutional level DRI support playing a stronger role in coordinating.

Requirement: Support the sustainability and continuity of existing UKRI supported AI capable systems, e.g. through grant extensions and/or recurrent funding for operating costs

Moving forward, it will be necessary to support greater coordination amongst the existing HPC facilities around their support for the development / adoption of AI with a view to supporting a more complementary and coordinated provision of compute. A national DRI strategy should support this. There is also a benefit to increased coordination and knowledge sharing at a more operational level. This should include greater coordination and collaboration around facilitating access to and movement of data, for example through standardised logins / access routes across a range of compute services, as well as increased connectivity and interoperability of compute systems as far as possible. This will entail agreement on the baseline support for various libraries and software, job submission rules and a unified web interface. Such efforts should aim to improve the efficient use of the current and planned future systems and lower the barrier to entry to such systems.

Such coordination and collaboration should also extend to the provision of training and support to their user communities. Though existing facilities need greater support to cover the operational expenditures, there are also opportunities for greater cross-facility collaborations in terms of training. The Digital Research Infrastructure Retreat[19] held in March 2022 marks a first step towards increasing collaboration and coordination between facilities and would benefit from further consolidation and formalisation moving forward.

By way of example, the German Association for National High-Performance Computing (NHR) was founded in 2021 as a collaboration between eight universities and institutes in Germany to provide mid-level compute resources (Tier 2). Funded by national and state governments, the NHR supports centres to combine and coordinate their activities in terms of application areas, methods and training, and delivering collaborative projects.

Requirement: Support increased coordination and collaboration among DRI providers, e.g. through initiatives such as DRI Retreats

The AI research community has different needs and ways of working than other data intensive research fields. Those working in **AI research and innovation often require greater interaction with the compute in real time** via tools such as Jupyter Notebook and RStudio. However, traditional HPC infrastructure has often been set up to support large scale batch jobs, which are ill-suited to the needs of AI researchers. To be able to support AI research and AI enabled interdisciplinary research, HPC clusters providing compute for AI should continue to explore the potential of incorporating cloud technology elements which would facilitate AI researchers' workflows.

Some existing Tier 2 facilities are already working to provide web-based accessible systems for their users, whilst others should be encouraged to do so. This could be aided by the provision of an open-source web-based supercomputer interface that could be reused on any system, such as the SAFE tool developed by EPCC. This will better enable HPC infrastructures to keep pace with the rapid developments in AI applications, tools and libraries and lower the barrier to access.

Requirement: Encourage development and uptake of tools improving the accessibility and consistency of DRI systems, e.g. notebooks and containers/virtualisation

Overall, interviews demonstrated fractured awareness of the various facilities and services available in the UK. This can in part be attributed to the complexity of the ecosystem and to the fact that many investments and programmes are relatively new. However, it is also indicative of a larger issue around the awareness of activities and facilities in DRI supported by other research councils. The DRI ecosystem would benefit from having an **overview of facilities and how to access them that is clearly available and easier for researchers to find at a UKRI-level**, rather than having to navigate multiple websites. Despite the existence of InfraPortal, interviewees and respondents to the survey expressed interest in a centralised directory / catalogue of existing infrastructures and resources available, perhaps indicative of a low awareness of the portal. In addition, there

[19] N8 Centre of Excellence in Computationally Intensive Research (2022) N8CIR Digital Research Infrastructure Retreat – https://n8cir.org.uk/dri-retreat/

were calls for collaboration with research groups at the institutional level to help raise awareness of resources available, perhaps through a network of ambassadors. As noted above, the use of such facilities, especially by communities working with virtual notebooks or new to the facility, depend on clear, up-to-date and readily available documentation of system architectures and parameters to ensure the system meets their requirements.

Requirement: Raise awareness of DRI for AI facilities, support availability and access models, e.g. through DRI directory and DRI ambassador network

Notably, the emergence of new and different hardware accelerators for AI may result in the diversification of the compute provision. Though GPUs were the first AI hardware accelerators and are now the most common, others such as Vision Processing Units, Field-Programmable Gate Arrays, Application-Specific Integrated Circuit, Intelligence Processing Units or Tensor Processing Units are gaining traction. These platforms are more specialised and the appropriateness of each depends on a wide range of parameters, (e.g. workload types, algorithms, memory and bandwidth requirements, etc.). The uptake of these alternative and more specialised platforms is unclear and over 60% of respondents indicated they didn't know how many other accelerators that are not GPUs they would need in five years' time for either their largest or typical workflows.

To further explore these technologies, research groups and infrastructures would benefit from specific funding projects or programmes to undertake technology foresight work to test and experiment with their capabilities, build technical knowledge required to run these systems as well as building a wider user base. Such a programme could also make a positive impact on supporting the development of UK based hardware companies through procurement, co-design and evaluation of future technologies. Some such technologies are already being tested under the scope of the ExCALIBUR programme. These testbeds could be helpfully made available to a wider community of potential users; however their grant funding is strictly time limited and will cease at the end of the programme. This foresight work should also be a long-term activity to enable the continued learning and development for the community for future generations of technology and researchers.

Requirement: Explore the potential of next generation AI systems through a technology foresight and horizon scanning initiative, e.g. building on ExCALIBUR testbed approach

The UK needs a more competitive high-end machine in order to not be left behind. As set out above, other countries already have Tier 1 class facilities with GPU accelerators, putting the UK at a competitive disadvantage. Such facilities offer capabilities beyond what the UK provision is currently able to support.

There is currently no large-scale national compute system for AI available to researchers in the UK. The UK's national supercomputing service ARCHER2 does not include the accelerator hardware required for most AI approaches.

Whilst scaling up the power of compute provision is (and will continue to be) necessary, any investment needs to be demonstrably transformative to the work of the user community. It was clear from the survey results that researchers who engaged with the review were generally working with small compute systems or resource allocations, with 82% of respondents using a maximum of 16 GPUs for typical workloads. 6% of our survey respondents indicated currently needing more than 513 GPUs for their largest workflow, while 17% estimated needing access to over 513 GPUs for their largest workflows in five years' time.

Interviewees noted that without access to and funding to support research on such a facility, the extent to which is realistically possible to envisage one's compute needs is going to be limited. However, by increasing both, the current compute capacity and researchers' supported capabilities, it is very likely to increase the scale of compute used by researchers and widen the community of researchers doing AI at scale. Moreover, the number of researchers working to develop or apply AI is also likely to increase significantly, especially in the light of other investments under the scope of the National AI Strategy.

Large-scale compute provides valuable resources for addressing societal challenges. Such computational provision provides a valuable resource for addressing global societal challenges. Most AI strategies globally currently focus on sectors with the highest potential for AI to have a transformational impact such as health care, mobility and transportation, agriculture and food, and the energy sectors.[20] Investments in compute should be made relative to the UK's policy objectives and the thematic areas / specific challenges to be addressed here in the UK. Identifying which thematic areas / use cases would be of specific priority / benefit most from such a facility would need a thematic focused approach.

A 2020 report from the Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative analysed global AI strategies and noted that governments currently focus on AI opportunities in health care, technology, agriculture, and manufacturing, with the rationale that these sectors have the highest potential for transformation through AI application.[21]

Examples of supercomputers addressing real-world challenges

The combination of supercomputers and AI have already proven to be effective in helping governments, as well as industry, in addressing the most complex issues and challenges, ranging from pandemics to climate change. For example:

The Summit supercomputer at Oak Ridge National Laboratory's computational power, together with its support for AI and data analytics tools was used for work fighting COVID-19. It allowed researchers to create an efficient drug discovery process (the work won a special Gordon Bell Prize for work fighting COVID-19, referred to as the Nobel Prize of supercomputing), as well as to train a BERT NLP model on an extreme scale molecule database that can speed the discovery of new drugs.

Japan's Fugaku supercomputer was used to develop an AI model to predict flooding from tsunamis in "near real-time". The model is based on early observed offshore tsunami waveforms and can be used for rapid evacuation notices and disaster preparation.

Europe's supercomputers and their AI capacity will be an essential part of the Destination Earth (DestinE) project aiming to develop a highly accurate digital model of the Earth to monitor and predict the interaction between natural phenomena and human activities and to help to build resilience to climate-change

Finding: AI is already proving invaluable in addressing societal challenges in key areas such as Sustainable Development and the response to the COVID-19 pandemic

Finding: Adoption of AI tools and techniques is rapidly accelerating and proliferating outside of core areas such as Computer Science

Comparator countries are already investing in the next generation of facilities. For example, in the United States, the Oak Ridge National Laboratory's Frontier supercomputer boasts 1.1 exaflops of performance and the Aurora, another exascale supercomputer planned for the Argonne National Laboratory also in the United States, will have around 2 exaflops capability. The potential AI use cases for these systems have not fully been explored yet, however they will provide unprecedented opportunities for future research and development, including AI. Big steps up in the scale of available compute can make new cases of problem tractable and unlock new types of approaches for solving them.

Though compute to support large scale jobs / processing very large datasets is only currently needed by a small part of the AI community, these opportunities have potential to grow, providing transformational opportunities to the UK. Such a national facility should therefore employ access models that support both large scale workloads that employ the full compute capacity and a collection of small to medium sized workloads running parallel.

The development of these opportunities will also require parallel support for skills development and research to build the foundation upon which such transformational research is conducted. Therefore, increasing investment in skills and research in the coming months and years will be necessary to realise the full potential of a national facility.

Requirement: Accelerate planning for a national Tier 1 scale facility by feeding in review findings and requirements, ensuring that it is internationally competitive

[20] Galindo, L., Perset, K. and Sheeka, F. (2021) An overview of national AI strategies and policies. OECD Going Digital Toolkit Note, No. 14. https://goingdigital.oecd.org/data/notes/No14_ToolkitNote_AIStrategies.pdf

[21] Fatima S., Desouza K.C., Dawson G.S. (2020) How different countries view artificial intelligence. Brookings Institution. https://www.brookings.edu/research/how-different-countries-view-artificial-intelligence/

3. Data access

3.1 Current use

The large-scale computing facilities in the UK are further supported by a data transfer, storage, and analysis infrastructure. This includes JASMIN – a storage and analysis platform provided by NERC and STFC for climate and earth sciences applications, Janet – a high-speed fibre-optic network for the academic community connected to GÉANT, the pan-European data network for the research and education community, and the UK Research Data Facility (RDF) – an EPSRC funded facility providing high-capacity disk and tape storage for data from national large-scale computing facilities.
Beyond these facilities exist a large and complex landscape of data infrastructures. The Open Data Institute (ODI) has been leading work to map and profile the landscape of data institutions in the UK, and is compiling a living register of data institutions from around the world.[22] As of April 2022, the UK has 89 individual organisations that could be classified as data institutions, responsible for facilitating access to data, combining, or linking data. Of these 35, institutions are responsible for publishing open data.

3.2 Barriers to access

The lack of availability of data for AI is a common problem across research and innovation communities and presents a barrier to almost all AI-related research fields. This is especially true for those research fields without a strong legacy of primary data collection and curation. Interviewees noted a range of challenges in accessing data, which varied according to the nature of the research being conducted and the data required.

The majority of survey respondents source their data from a combination of open / freely available data sources (77%), academic collaborators (69%), or their own sources (61%). The majority then stored this data on institutional / organisational services (89%) or on their individual computers (70%).

One-third (35%) of respondents indicated that **data owners being reluctant to share private / commercial data was a significant barrier** for the availability of data for AI. This reluctance can be due to a range of different reasons: from commercial concerns around IP protection, to legal impediment and legal uncertainty, to ethical and regulatory issues, to a lack of awareness of the opportunity or the poor / unstructured nature of the data itself. For a researcher, accessing such data also comes with costs associated with licensing and legal fees, as well as the time and effort required to negotiate access and ensure compliance with regulations.

Finding: In spite of open data initiatives from UKRI and other public bodies, the potential of AI for research and innovation is not being fully realised due to issues around data access

Finding: Commercial datasets can be prohibitively expensive for researchers to license

Around a third of survey respondents also indicated that the time required to adapt existing data for AI purposes was an important barrier in relation to the availability and suitability of data for AI. However, interviewees also noted that the time required to adapt data for AI purposes was often an inherent aspect of conducting AI research and often a valuable process for understanding the context, format, and potential limitations of a dataset.

It is often a **challenge to work with data from multiple sources because of access models and lack of data interoperability**. Interviewees attributed this to siloed working between research organisations, data providers, and disciplines, which have limited awareness and sharing of data between communities and therefore limited cross-disciplinary research. Additionally, those wishing to use and access data from multiple environments face hurdles in completing multiple processes for securing permissions, and then for linking data sets. This is reflected also in the experiences of the survey respondents, 45% of whom indicated the need to combine data from multiple data sources as either a significant or moderate barrier.

Finding: Data interoperability is becoming increasingly crucial to support use of AI techniques in an interdisciplinary research context

As it stands, public sector data is of varied quality and often difficult to access. There is a wealth of data collected and held by government and the public sector that is of value to the private and third sectors. However, interviewees highlighted that data was of variable quality, and often not available in a usable and consistent format. Subsequently, it is also more challenging to identify opportunities for improvements in accuracy, efficiency, and accountability of public policies.

Finding: Researchers advise that broader adoption of AI will require significant effort around standards and processes for collating and organising data

Data storage and sharing may also be subject to wider ethical, legal, and political complexities. The extent to which researchers and innovators are able to obtain or use data can depend heavily upon legal requirements and licence agreements around the storage and sharing of data. These laws can often be country specific, creating challenges when working internationally, or sector specific, requiring specific knowledge and qualification to secure access or publish. In addition, the governance structures surrounding sensitive data are in themselves complex, with multiple stakeholders from across the public, private, and third sectors. As a result, researchers are often required to contend with a range of ethical, legal, and operational challenges to conduct their research. As these challenges most often pertain to sensitive data in areas with significant potential for wider societal impact, they merit focused activities to overcome.

Finding: Researchers advise that concerns about potential ethical, legal, and political complexities can have a significant effect on data sharing and re-use

3.3 Future needs

As noted above, gaps in data availability are presenting barriers to AI research and innovation across almost all research fields. Those consulted for this review also indicated their data needs were often not only research field specific but related to their specific research questions and challenges and the data they are looking to work with. As a result, identifying specific challenges relating to data for AI in particular research fields is challenging and demands a granularity of data collection and analysis beyond the scope of this review. To identify specific data gaps in AI-related research fields, UKRI and other stakeholders will likely benefit from facilitating community led processes to identify and agree upon key priority datasets. The future needs set out below are those common across different research fields.

The amount of data that researchers are working with is set to increase significantly in the coming years. The size of datasets that survey respondents were working with covers a wide range, between 1 gigabyte and 100 petabytes (1 petabyte = 1 million gigabytes) of storage capacity. On average, respondents currently require around 1 terabyte (1 terabyte = 1 thousand gigabytes) of working storage for their largest workflows and up to 100 terabytes of overall storage capacity. On average, respondents to our survey expect that in five years' time their largest workflows will require around 10 times more working storage, approx. 10 terabytes, and around 10 times more of overall storage capacity, up to 1 petabyte. Several responses indicating that some of the largest workflows are expected to be of size of 100 petabytes or even reaching 1 exabyte (1 exabyte = 1 billion gigabytes). This is considerably larger than the capacity of both current systems provided in the UK, and many future systems planned internationally. For example, the Oak Ridge National Laboratory's Frontier supercomputer in the United States is expected to have 1 exabyte of storage capacity.

Finding: The amount of data that researchers are working with is expected to increase tenfold over the next five years

Globally, other national plans for AI also recognised that facilitating access to data was among the most expressed outcomes for national AI plans and strategies.[23]

Broader adoption of AI will depend on the development and implementation of standards

[22] Open Data Institute (2021) The Data Institutions Register – https://theodi.org/article/the-data-institutions-register/

[23] Fatima S., et al (2021) Analyzing artificial intelligence plans in 34 countries. Brookings Institution – https://www.brookings.edu/blog/techtank/2021/05/13/analyzing-artificial-intelligence-plans-in-34-countries/

and processes for collating and organising data for AI in line with FAIR principals. Some research fields benefit from a legacy of collaborative working and the development of international standards and processes for managing data, whilst others do not. Interviewees indicated that such efforts should be community led and driven by researchers to ensure that such standards reflect their requirements. To support this, specific funding and projects will be needed to support research communities to develop data management standards and communities of practice in research fields where data-intensive research is emergent.

This is especially true for research fields in which AI is nascent. For example, interviewees highlighted projects and programmes such as the Physical Sciences Data Infrastructure[24] and Living with Machines[25] as providing valuable hubs to facilitate the development of standards and practices for sharing data between institutions and communities of practice, thereby supporting the interoperability of data. Notably, however these projects are time-limited without a clear path for sustained support.

In all cases, considerations should also be made for domain specific requirements for data standards to enable AI. The implementation of agreed sets of standards and best principals to create FAIR and AI-ready experimental data could create the accessible and efficient data foundation required to produce novel AI tools and enable discoveries in science, technology and engineering, as well as evidence to inform policies.

There have been few attempts to develop data standards for the AI community. One of the more well-known initiatives is also a tool, called Datasheets for Datasets[26], which is designed to standardise the process of documenting the datasets used for training and evaluating machine learning models. The tool aims to facilitate better communication between dataset creators and those using datasets to train machine learning models, and encourage the machine learning community to prioritise transparency and accountability. According to the authors, the tool can benefit both groups (creators and consumers):

> *"For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use. For dataset consumers, the primary objective is to ensure they have the information they need to make informed decisions about using a dataset".*

Since 2018, when the original article was published, the work has gained traction not only in the academic setting but also industry. IBM and Google have followed this work with their own takes on datasheets, named FactSheets and Data Cards respectively.

In the longer-term, wider development and adoption of AI would be supported by widespread documentation of research datasheets. To support this, there is a need to encourage and incentivise the widespread adoption of data standards necessary for AI in association with open data policies. As noted in the Final Report of the Open Research Data Task Force, there is also a need to ensure availability of appropriate funding to enable the development and maintenance of open research data through direct funding, research project grants and through other routes such as Data Study Groups.[27] As such, documentation of research data could be made a requirement of UKRI grant awards. It is also important to note that not all datasets can be shared, but even restricted datasets can be made more easily accessible to others by adopting data standards.

Requirement: Support development and adoption of data standards to ensure research data is AI-ready

There is also a **need to support the interoperability of data to support interdisciplinary AI research**. Siloed working between research organisations and disciplines has limited the degree of awareness and sharing of data between communities and therefore limited cross-disciplinary research.

To address this, there is a need to explore how and where some core requirements for how data is stored and metadata used could be implemented across disciplines. Interviewees who conduct interdisciplinary research also noted the challenges in using and accessing data from multiple environments, including the multiple processes for securing permissions and then for linking datasets. This work should be tied into UKRI's open data requirements, perhaps through the provision of guidance from UKRI for research organisations to incorporate into institutional level research data management roadmaps.

Requirement: Support for interdisciplinary AI research, e.g. by identifying core data storage and management requirements across disciplines

Further support and consideration are needed for the curation and maintenance of existing large unique public and scientific datasets. These provide a valuable resource across research fields but require both capital and operational expenditure to produce and maintain.

Some experimental datasets are not in a suitable format to fully exploit data-driven discovery. There is therefore a need for funding to support the development of FAIR datasets and models for AI research and innovation that are reusable. Programmes like the US Department of Energy's FAIR Data and Models for Artificial Intelligence and Machine Learning funding call provide focussed support for researchers to make publicly released datasets and models comply with the FAIR principles, and to provide guidance to other researchers on how to do the same. The UK should explore providing a similar support to its researchers too.

As it stands, **public sector data is of varied quality and often difficult to access**. There is a wealth of data collected and held by government and the public sector that is of value to the academic, private and third sectors. However, interviewees highlighted that data was of variable quality, and often not available in a usable and consistent format. In many fields, this would primarily entail enriching existing datasets rather than creating or releasing new ones, focussing on improving their quality, consistency, and interoperability for AI.

Finding: Public sector data can be of varied quality and can be difficult to access

Requirement: Improve accessibility and quality of public datasets for AI, e.g. by supporting development of exemplar datasets and supporting tools/documentation

Facilitating sharing of commercial / private sector data will be of critical value to a broad range of research fields and industries. However, the mechanisms and approaches for doing so will need to be tailored to specific sector requirements or nuances and delivered in collaboration with key stakeholders. Overall, however, UKRI and the Government should continue developing initiatives and policies to improve access to and sharing of private sector data, e.g. through exploring data trusts or data cooperatives.

Certain research fields would benefit from a centralised organisation that is responsible for collating, standardising and / or integrating datasets from disparate sources. Though these datasets may become outdated quickly, they also have the potential to provide valuable historical data for research communities. Such organisations can also play a key role in reviewing and critically evaluating datasets (e.g. for bias, gaps, or more inherent structural issues), which is especially valuable for researchers with less experience working with large data sets.

Alternatively, centralised organisations or groups can play an important role to mediate or broker access to privately held datasets. Facilitating access to a dataset and supporting research groups or individuals to work through licence requirements for example.

Finding: Varied access and licensing models and data interoperability issues can make it challenging to combine data from multiple sources

Requirement: Explore mechanisms and instruments to support the collating of datasets from disparate sources and brokerage for access to commercial datasets

Datasets required for training AI models can be difficult for researchers to find and access, and are often massive and held separately from compute facilities. These factors combine to make adoption of AI techniques unnecessarily difficult.

[24] Physical Sciences Digital Infrastructure project website – https://www.psdi.ac.uk/
[25] Living with Machines project website – https://livingwithmachines.ac.uk/
[26] Gebru, T. and Morgenstern, J. and Vecchione, B. et al. (2018) Datasheets for Datasets. arXiv. https://arxiv.org/abs/1803.09010
[27] Open Research Data Task Force (2018) Realising the potential: Final report of the Open Research Data Task Force – https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/775006/Realising-the-potential-ORDTF-July-2018.pdf

As applications for AI emerge, there is a growing need to strengthen the compute provision associated with such centralised data sources. The approach to providing access to this compute depends on the relative size of the data, its sensitivity, and the compute demands currently and in the near future. Short-term investment in compute for AI should explore the potential of co-locating key datasets and AI compute facilities on high performance storage, as this is required to fully realise the benefits of the investment in accelerator hardware. The model and approach for doing so will need to reflect the particular nature of the data, the location of the data and the needs of the users, with consideration for privacy and licensing concerns.

Of note, the Data and Analytics Research Environments UK (DARE-UK) initiative is investigating Digital Research Infrastructure requirements for Trusted Research Environments (TREs) working with sensitive information such as administrative records and medical data. The review notes that there is significant researcher interest in provision of AI capabilities as part of these environments. DARE-UK has already supported work in this area and is strongly encouraged to continue investigating how this may best be delivered.

Requirement: Explore the potential of co-locating compute and data for key large scale public datasets and Trusted Research Environments

Interviewees noted that the need for greater compute provision would require investment in compute and data facilities but was also critically dependent on support from Research Technology Professionals with the appropriate skills and expertise. This aspect is discussed in the next section.

4. People and skills

Almost all interviewees agreed that any investment in any infrastructure for AI would need to be matched by investments in training and support.

Supporting this, survey respondents indicated that the three highest priority areas to meet their current and future needs (after access to computing systems with GPUs) were funding for Research Software Engineers (62%), training for researchers (61%), and funding for general technical support services (61%).

In this section, the findings and key requirements for training and support in relation to key groups or communities, including those responsible for running DRI, domain specialists with AI skills, AI specialists, and the wider research and innovation community are presented.

4.1 DRI research operations

The survey and interviews emphasised the importance of having access to Research Technology Professionals (RTPs) within institutions or departments. Researchers need expert support to help them with adopting AI tools and libraries and best development practices, as well as exploring and exploiting DRI for their research. In particularly this has been emphasised in disciplines where AI adoption is still in its infancy. As these teams and staff are often the first contact point for students and researchers with nascent computing needs, there is a need to ensure they are well staffed and resourced to be in a position to provide sufficient support.

However, interviewees noted that the teams of research technology professionals within universities are often relatively small and working to support the breadth of needs across the university. Institutions and providers of DRI also often struggle to recruit individuals with specialist skills to manage institutional and Tier 2 level facilities. Publicly funded institutes are not in a position to offer salaries that are competitive with industry, or at times even permanent job positions, and these teams are often small and expected to collaborate and support a large and increasingly diverse community of researchers.

Finding: Teams of Research Technology Professionals within universities are often relatively small and working to support a breadth of needs across the institution

Finding: Central funding for DRI has often been capitalised, with little or no support for staffing to assist researchers in adopting and exploiting the infrastructure

Improving the pay structures and career paths for RTPs to increase competitiveness with private sector will go some way towards addressing the challenges around recruitment. In addition, complementary investments to develop and promote career development pathways around AI and data science, such as Machine Learning Operations (MLOps), as a key component of Research Software Engineering would be beneficial. In the longer-term, this could evolve and formalise to a specific AI Research Technology Professional career path.

There is a need to provide greater support to these groups and ensure that career paths and pay structures reflect the importance of these individuals within the university and are more competitive with industry.

Funding models for many public DRI facilities are complex and involve multiple streams of funds with varied duration. Usually capital expenditure is well supported, however, the day-to-day expenses of running and supporting these facilities often rely on less established resources that depend on hosting sites, partner organisations or commercial partners.

63% of survey respondents prioritised access to RTPs within their research groups or institutions/organisations. Interviewees indicated that the resource allocation for such teams within the university can be tied to funding from specific research departments or groups or tied to specific research projects. In addition, where RTPs are included within research grants, they are subjected to both institutional and funder level requirements and limitations. This can leave less time resource to support students or research groups in fields with a nascent or emerging use of AI or data intensive research. There is a need to increase the core / base level of funding to academic research computing teams within institutions to support a wider community of research fields.

Finding: DRI providers can struggle to recruit and retain staff with the necessary skill sets due to issues around contract length, pay scales and progression opportunities

Finding: Staffing may be constrained to particular projects, e.g. where a project has obtained funding for Research Software Engineering support

Requirement: Explore potential funding models to support RTP career paths, competitive pay structures and job security within DRI facilities and institutional teams

Requirement: Explore approaches to providing dedicated institutional DRI staff capacity

As future exascale systems will support both traditional modelling and simulation applications, as well as large-scale AI applications / workflows, there is a need to ensure the research technology professionals are equipped with the requisite skills to support this. The breadth of research fields adopting compute is widening, so teams based within institutions and DRI infrastructure need a widening set of knowledge and skills to meet these needs. Moreover, as the hardware and software tools being used by the community are constantly evolving, and particularly quickly within AI, there is a need to keep pace with these developments.

To ensure the skills are in place to operate this infrastructure there is a need for a breadth of scales and types of training programmes. DRI staff would most benefit from modular programmes or courses, allowing them to focus their learning to reflect the needs of their respective facilities, user communities, and existing skills levels. These courses would most helpfully be embedded within other training programmes designed for the RSE community. The Research Software Engineer Knowledge Integration Landscape Review also identified a need for dedicated training programme for RSEs who want to focus on HPC and a long-term training and education strategy to ensure gaps in training and knowledge are addressed, including those relating to AI.[28] A good example of current initiatives targeting more general RTP upskilling are the training courses organised by the Software Sustainability Institute and ARCHER2. Similar initiatives are required for developing AI operations skills.

The key skills identified by ExCALIBUR's RSE landscape review as being required by the AI / HPC community include the ability to understand surrogate models, containerisation, scalable AI algorithms, algorithms to quantify uncertainty, HPC/AI hybrid application development, large-scale complex generative models, debugging and profiling AI models, modification of open-source frameworks, and data lifecycle management. In addition, there is a growing need for all individuals working with and supporting AI development and implementation to be familiar with principals and approaches for ethical and responsible AI.

These training programmes and documentation should be continually maintained and well-curated, as well as being easily accessible to individuals working with DRI facilities across the landscape. Specific fellowships and programmes that focus on the practical development and operation of AI within infrastructures, research groups, and companies may also support this.

Requirement: Provide opportunities for RTP staff upskilling in AI, e.g. through a programme of training and supporting resources, focusing on AI

4.2 Domain specialists with AI skills

The wider value and impact of AI to the research community will emerge from interdisciplinary collaborations. To enable this, there is a need for cross-domain specialists with expertise in AI who are also able to work collaboratively with domain specific researchers to support the application of AI tools to their workflows. As it stands, many research communities only have a limited number of individuals who can "translate" the different needs and requirements from an AI perspective and a domain specific perspective. This is especially valuable in research fields without a strong history of data intensive research such as in the arts and humanities.

To address this, there is a need for training programmes in AI skills with specific research field or sector focus. The Government's new Masters AI conversion courses will go some way towards addressing this need, however there would also be a benefit to longer programmes of vocational training courses available to ensure continued learning. Such training activities would most helpfully be embedded within existing programmes or initiatives to support data intensive research.

The review found that researchers from fields without a strong history of quantitative research (e.g. arts and humanities, and social sciences) often find it more challenging to engage with compute facilities and require more support and training to successfully access and use such systems.

Finding: Data owners and users can struggle to prepare their datasets for processing by AI tools due to lack of specialist expertise

Requirement: Promote uptake of AI techniques in under-represented disciplines, e.g. through support for AI training courses with domain/research field focus

4.3 AI Specialists

There is expected to be continuing and growing need for AI specialists, both in research and industry. Meeting this need will require support for PhDs in AI, but more importantly support for training for researchers and industry professionals. Self-directed learning is common in AI and most firms with employees in AI roles undergo informal or on-the-job training throughout their roles.[29]

However, the rapid evolution of AI and the scale of growth will require a broad set of upskilling programmes and initiatives from actors across the DRI ecosystem, targeted at a range of different knowledge and experience levels, and focussed on different aspects of skills needs.

When asked in the survey which skills respondents would like to prioritise to develop or further improve in order to improve and maximise their use of DRI: 60% prioritised machine learning frameworks (such as PyTorch and TensorFlow); 44% – Data analysis and Parallel/accelerator programming and/or distributed learning; 34% – Best practices on software development/coding and Organising and structuring data and/or code.

Finding: Fast paced development of AI tools and techniques highlight the need for truly continuous and ongoing professional development

Requirement: Continue to support training of AI specialists to maximise their use of DRI, e.g. by building on initiatives such as Turing AI Fellowships and Centres for Doctoral Training in AI

4.4 Wider research and innovation communities

More broadly, interviewees noted the importance of building operational knowledge of AI within industry to support adoption of AI. In particular, building experience and knowledge of the risks and overall mechanisms for AI will be important for supporting industry engagement at the more senior levels.

Similarly, interviewees highlighted a need to support an increase in the basic knowledge / awareness of AI within research groups. Though they do not necessarily need to become experts, senior lecturers or supervisors in academia should be trained in AI / coding to a sufficient level to be able to train / oversee the training and adoption of AI in research undertaken by PhDs / PDRAs.

Requirement: Engage with industry and academia to raise awareness of the potential of AI, e.g. through upskilling and training programmes

A key aspect of supporting access to and use of DRI for AI relates to the development of communities of practice within research communities. Such communities are essential to facilitating cross-pollination of expertise, sharing tools to minimise duplication of effort, and enabling new AI researchers to access AI scientific software. Such communities also **help to address data availability, where AI researchers are working directly with the collectors / curators of data**. The groups also work to build a national community of users that are connected to DRI, aware of the capacity, and informed about access.

The study sought to profile the needs of

[28] ExCALIBUR Project (2021) Research Software Engineer Knowledge Integration Landscape Review – https://excalibur.ac.uk/resources/research-software-engineer-rse-knowledge-integration-landscape-review/

[29] Ipsos Mori (2021) Understanding the AI labour market: 2020 – https://www.gov.uk/government/publications/understanding-the-uk-ai-labour-market-2020

researchers who are interested in using AI in future, but who are not currently doing so. However, engaging with such individuals proved challenging and the survey secured only c.15 responses from this group. As a result, the extent to which their perspectives are presented within the study is likely limited. However, interviews with stakeholders suggested that to support the adoption of AI amongst the wider research community, a breadth of communication and engagement activities would be necessary to demonstrate the tools and applications of AI within their research field. There was also felt to be a need to increase awareness of access to support and training of relevance to their research field, as well as to provide mechanisms to bring together AI researchers with those new to the field. To support this, existing DRI's providing access to AI infrastructure would benefit from small scale outreach / support grants to enable such work. Ideally, this support would be complemented by a small-scale cross-DRI programme to share best practices and resources and further strengthen the connections amongst the network.

Requirement: Support for communities of practice and interdisciplinary collaboration, e.g. through demonstrators showing the potential of AI in nascent fields

Interviewees often noted that the application of AI tools for research and use of DRI for AI within AI nascent fields and sectors often depends on the wider skills base within the research and innovation community. Even amongst researchers currently developing or applying AI, 37% of respondents reported they currently had poor or very poor skills in organising and structuring data and/or code.

For those interested in using AI as a research tool, whilst there is no need to become specialist in AI, a basic understanding of the core principles is still necessary to support effective collaboration and implementation of AI.

Increasing the foundational awareness and familiarity with data science methods across research fields was thought to be a valuable step towards increasing the pipeline of researchers and professionals equipped to engage with AI in the longer-term. Addressing this will require a wide set of programmes and initiatives across the research and innovation landscape in the UK, such as embedding coding and data science methods in undergraduate degrees, industry or sector specific upskilling programmes, increasing access to informal AI training courses, etc.

The full breadth of the training necessary is beyond the scope of DRIs to provide, however they could helpfully provide training courses or guidance tailored to non-expert users to support the use of their respective facilities.

Finding: There are significant gaps in training and knowledge – 37% of survey respondents said they had poor or very poor skills in organising and structuring data and/or code

Finding: Training and documentation often do not reflect the differing needs of AI researchers ("tool builders") and the wider research community ("tool users")

Finding: Computer Science, Physics and Engineering were particularly strongly represented (two thirds of survey responses) – further work may be desirable to engage with the wider research community

Requirement: Support for DRI providers to engage with non-expert users, e.g. through training, guidance and resources

5. Overall conclusions

Digital Research Infrastructure (DRI) provides a key resource for addressing current and future societal challenges, such as health, climate change, food security, and sustainable energy. Addressing such challenges will inherently require a breadth of funding for increased AI compute capacity and support to provide the skills and resources to maximise the opportunities such infrastructure provides.

The UK needs to continuously invest in DRI for AI if it seeks to realise its ambition of being a global AI superpower

Demand for compute and data for AI has grown significantly in recent years and is expected to continue to do so. In order to address the future needs of DRI for AI, activities, investments and programmes will need to support a scaling up and scaling out of compute provision, increased consolidation of data, the operational running of both compute and data facilities, and the necessary training to support uptake and sustainability. Failing to invest in DRI for AI, the UK will not be able to support many new activities and will weaken its position against the ambition of becoming an AI superpower.

Compute for AI needs to be accessible, configurable, adjustable, and promote collaboration

Demarcating compute provision using a tiered approach overall seems to work well, providing researchers with different scales, architectures and levels of support to meet different needs. However, the current systems available in the UK within Tiers 1, 2 and 3 are either working at maximum capacity, under-resourced, limited in their compute provision specifically to support AI, or a combination of all three.

The vast majority of researchers expect their compute needs for AI research and innovation to more than double in five years time and indicated that access to computing systems with Graphics Processing Unit (GPU) accelerators would be a priority. However, the breadth of researchers and needs requires a breath of compute resources at different levels.

Tier 1: An internationally competitive large scale national facility

There is currently no large-scale national compute system for AI available to the UK's research community. The UK's national supercomputing service does not include the accelerator hardware required for most AI approaches. As a result, the UK is at a competitive disadvantage compared to other countries who now provide this capability in their own flagship services, as well as those countries with access to EU-level initiatives. Comparator countries are also already investing in the next generation of facilities which will provide the next level of AI compute capability, which will further widen the compute capacity gap with the UK.

Accelerator hardware, such as GPUs, to provide AI capability should be incorporated into the UK's next national scale compute service as this is now required for general use by the compute intensive research community. These new machines will allow for new technologies and research via the power of computing and simulation methods, data analytics and AI, and combination of all, to address societal challenges.

However, such a facility requires a higher level of investment and would take some time to establish. As there is an immediate need to increase the compute provision for researchers, there is a need to look to potential solutions in the near-term.

Tier 2: Uplift existing Tier 2 facilities

The existing network of Tier 2 facilities provide valuable regional hubs of expertise distributed across the UK, each working with different user communities. However, the roll-out of AI hardware somewhat varies across the existing Tier 2 facilities, which are also often currently working at maximum capacity.

In the near-term, there is a need to increase the available compute capacity at Tier 2 facilities, as well as the associated support capacity (both in terms of staff numbers and capability) in order to support their current and potential wider AI user communities. In association with this, Tier 2 facilities should continue to adopt cloud type approaches such as notebooks and container-based virtualisation, which lower the barriers of using compute for AI and better enable researchers to move between facilities.

In the short-term, there would be benefit to formalising, with associated resource, a mechanism for Tier 2 facilities to coordinate and collaborate to ensure efficient and coherent training and support, both internally and externally, and coordinated delivery of hardware technologies. In the medium-term, the continued support of Tier 2 facilities provides a valuable tool for the phased development and provision of compute for the AI and wider research communities.

Tier 3: Encourage uplift of institutional compute provision and support

As institutional and research group level sources of compute are currently researcher's primary sources of compute, there is a need to encourage the uplift and support of these resources. These university level teams can also provide valuable "on the ground" support for researchers and a first port of call for students / individuals looking to explore AI for their research.

In the medium-term however, there is a risk that Tier 3 level investments further exacerbate existing barriers around interoperability of systems and data. Therefore, there remains a need for a mechanism to support a coherent strategy and interoperability with Tier 2 and Tier 1 level systems.

Embed AI needs in a coherent and coordinated DRI roadmap

The current UK AI DRI ecosystem has evolved over many years, rather than being "designed", to support a diverse range of communities through numerous funding sources and mechanisms. As a result, it is complex and somewhat fragmented. To work towards a more coherent and coordinated DRI ecosystem for AI, there is need for a single DRI roadmap that sits across all constituent UKRI Research Councils and is embedded within a wider national DRI strategy. This roadmap should also cover the use of the commercial cloud, as well as the investment in exploratory co-design and evaluation of future technologies. This would provide clearer direction for long-term planning and include a framework and financial plan for longer-term investment for its maintenance and continual renewal.

Any investment in DRI must integrate with national priorities and overarching government strategies. This roadmap should be linked to other relevant strategies and roadmaps, such as the National Data Strategy and National AI Strategy. To meet the UKRI objective to deliver carbon neutral digital research infrastructure by 2040, any future investments in DRI will also need to carefully consider questions around resource efficiency and its carbon footprint.

Such a roadmap, and the associated integration of the national DRI system, would help lower barriers to accessing compute resources and data in a timely manner and support greater interdisciplinary knowledge sharing and research. Overall, the objective should be to support the development of a federated compute infrastructure that combines new and existing resources within a hybrid approach (at Tier 1, Tier 2, Tier 3 and commercial cloud) to connect users to a range of diverse systems. This infrastructure should also facilitate their use through the provision of educational tools and user support to enable their research.

There would be benefit to a single organisation playing a convening and coordinating role to deliver this roadmap, as it could provide a "single voice" and central contact point for the AI community.

Funding and programmes to support the adoption of AI largely sit at a Research Council-level, some of which have not historically made significant investments in e-infrastructure or computing. Additionally, some societal challenges may merit specific new infrastructures, whether they provide compute, data, or a blend of both. There is a need to support both scaling up (i.e. investing in a smaller number of larger AI systems) and scaling out (i.e. increasing provision across institutions) of compute provision for AI.

Any investment in hardware/compute for AI needs to be matched by investment in training and support

One of the strongest findings emerging from the Review was the need for training and support programmes for the full breadth of AI researchers and DRI staff. Most respondents indicated that after access to computing systems, the three highest priority areas to meet their current and future needs were funding for Research Software Engineers, training for researchers, and funding for general technical support services.

Researchers need expert support to help them adopt AI tools and techniques, particularly in disciplines where AI adoption is still in its infancy. In order to maximise the use of DRIs and to support researchers who do not have specialist skills to exploit large scale DRI facilities, funding will be required for upskilling and retaining professionals who operate and support DRIs. Funding is also required for establishing initiatives for developing AI operations skills for researchers as well as AI training courses especially with a domain / research field focus. Significant central funding has been provided for Centres of Doctoral Training in AI, however complementary central investments in a programme requiring greater coordination are needed in training and upskilling the Research Technology Professionals who operate and support DRIs, and to develop and promote career development pathways around AI such as Machine Learning Operations (MLOps) and data science as a key component of Research Software Engineering.

As it stands, most Tier 3 and Tier 2 level compute systems have limited capacity to support new users, develop and maintain training resources and tools, or undertake outreach activities. Therefore, there is a need for increasing the operational resource within DRI facilities to meet user needs. This is especially important for increasing DRI use by AI's nascent communities.

Making the best use of AI compute also requires a national community of users that are connected, aware of the capacity, and informed about access. Such communities provide a nexus to demonstrate the potential opportunities for AI, provide training and upskilling of a breadth of researchers, and support interdisciplinary collaboration. These communities can also increase data sharing, standardisation of data management and sharing practices, and facilitate access to and use of DRI, increasing awareness of the different facilities available and modes of access.

Unified data management standards and sharing policies are needed

The amount of data that researchers are working with is set to increase significantly in the coming years. However, this data is often challenging to access and not always suitable for AI. The adoption of AI will depend on the development and implementation of standards and processes for collating and organising and sharing data for AI in line with FAIR principals.

To support this, specific funding and projects will be needed to support research communities to develop data management standards and communities of practice in research fields where data-intensive research is emergent. There is also a need to explore how and where some core requirements for how data is stored and metadata used could be implemented across disciplines, which would support and enable interdisciplinary research and support the integration of data from disparate sources.

In the longer-term, wider development and adoption of AI would be supported by widespread accessibility and documentation of research datasets. To support this, there is a need to encourage and incentivise the widespread adoption of data standards necessary for AI in association with open data policies, for example as requirements of UKRI grant awards.

Increased industry engagement with DRI will require focused and targeted support, plus investment in additional compute capacity

As it stands, industry engagement with publicly funded DRI is relatively low. Instead, companies more often secure compute through other means, both directly and indirectly.

Any future investments in DRI for AI will need to include careful consideration for how such platforms or programmes would engage with industry partners, with specific operational models to meet specific industry needs. For example, industry partners often have greater concerns for data security and privacy, and the access models currently employed to secure time on public funded research infrastructure are often opaque. These organisations often also need other forms of business support and training, alongside access to such resources. However, as many compute facilities are currently working at or close to maximum capacity, any actions to increase industry engagement should be coupled with investments in additional compute capacity.

The extent to which this review has been able to capture and present the full breadth of industry needs from DRI in detail is limited and will merit further investigation. The review has explored industry engagement with DRI around AI use cases, however it has proved difficult to establish industry requirements. Further work needs to be undertaken to investigate industry requirements from a commercial perspective, taking a different and more targeted approach. This approach could helpfully focus more on engaging more with those industry networking organisations and trade associations that focus on supporting the AI sector, and those with sector specific foci.

5-10 year outlook

If implemented in full, the recommendations put forward by the community, as identified in this review, could amount to an integrated and holistic programme of support for compute capacity, data access, and people and skills. This would likely have an important impact on the UK's ambitions to be world-leading in AI research and innovation over the next 5 to 10 years.

The key benefits envisaged include more straightforward and equitable access to significantly enhanced compute capability for AI research and innovation, supporting a wider diversity of research communities, organisations, and geographic locations. The enhanced AI capability would incorporate cloud native technology where appropriate, and be complemented by a breadth of high-quality AI-ready open and public data sources. Improved arrangements would also be put into place for access to public sector data, restricted data and commercially licensed data.

In parallel, adoption of AI tools and techniques would be supported across research disciplines and in industrial R&D by developing and nurturing a highly skilled cadre of Research Technology Professionals and upskilling the wider research community. This would enable AI researchers to exploit DRI for AI to its fullest potential, through continued professional development, training opportunities and embedded support at an institutional level.

Acknowledgements

The authors would also like to thank all the stakeholders who gave of their time generously to input to the study through interviews, workshops, and correspondence.

This project was governed by a steering group of representatives from research councils and departments of the government:

Dr Liam Boyle, Engineering and Physical Sciences Research Council

Dr James Dracott, Engineering and Physical Sciences Research Council

Dr Victoria Mico Egae, Engineering and Physical Sciences Research Council

Oliver Grant, Department for Digital, Culture, Media & Sport

Dr Richard Gunn, Engineering and Physical Sciences Research Council

Dr Justin O'Byrne, Science and Technology Facilities Council

Prof Mark Parsons, Engineering and Physical Sciences Research Council

Yohann Ralle, Department for Business, Energy & Industrial Strategy

The authors are also grateful for the valuable time and input from the advisory group comprised of:

David Beavan, The Alan Turing Institute

Dr Nicolas Guernion, The Alan Turing Institute

Dr Steve Harris, National Health Service

Professor James Hetherington, University College London

Dr Scott Hosking, British Antarctic Survey

Dr Watjana Lilaonitkul, University College London

Dr Jim Madge, The Alan Turing Institute

Amit Mulji, The Alan Turing Institute

Professor Iain Styles, University of Birmingham

Nell Swanepoel, University College London

Dr Jeyan Thiyagalingam, Science and Technology Facilities Council

Organisations interviewed as part of the study

Airbus

Danu Robotics

Digital Catapult

Durham University

Earlham Institute

EMBL-EBI

EPCC

ESRC

Global Surface Intelligence

Hartree Centre

HPE

Intel

Kiwi Biosciences

Lancaster University

McKinsey

Met Office

Newcastle University

Quansight Labs

Queen Mary University of London

Queen's University Belfast

Relation Therapeutics

Roche

Rosalind Franklin Institute

STFC

Swansea University

The Alan Turing Institute

UK AEA

Unilever

University College London

University of Birmingham

University of Bristol

University of Cambridge

University of Edinburgh

University of Glasgow

University of Lancaster

University of Leeds

University of Leicester

University of Liverpool

University of Manchester

University of Oxford

University of Sheffield

University of Southampton

University of Sussex

University of Warwick

The Alan Turing Institute

technopolis group

turing.ac.uk
@turinginst