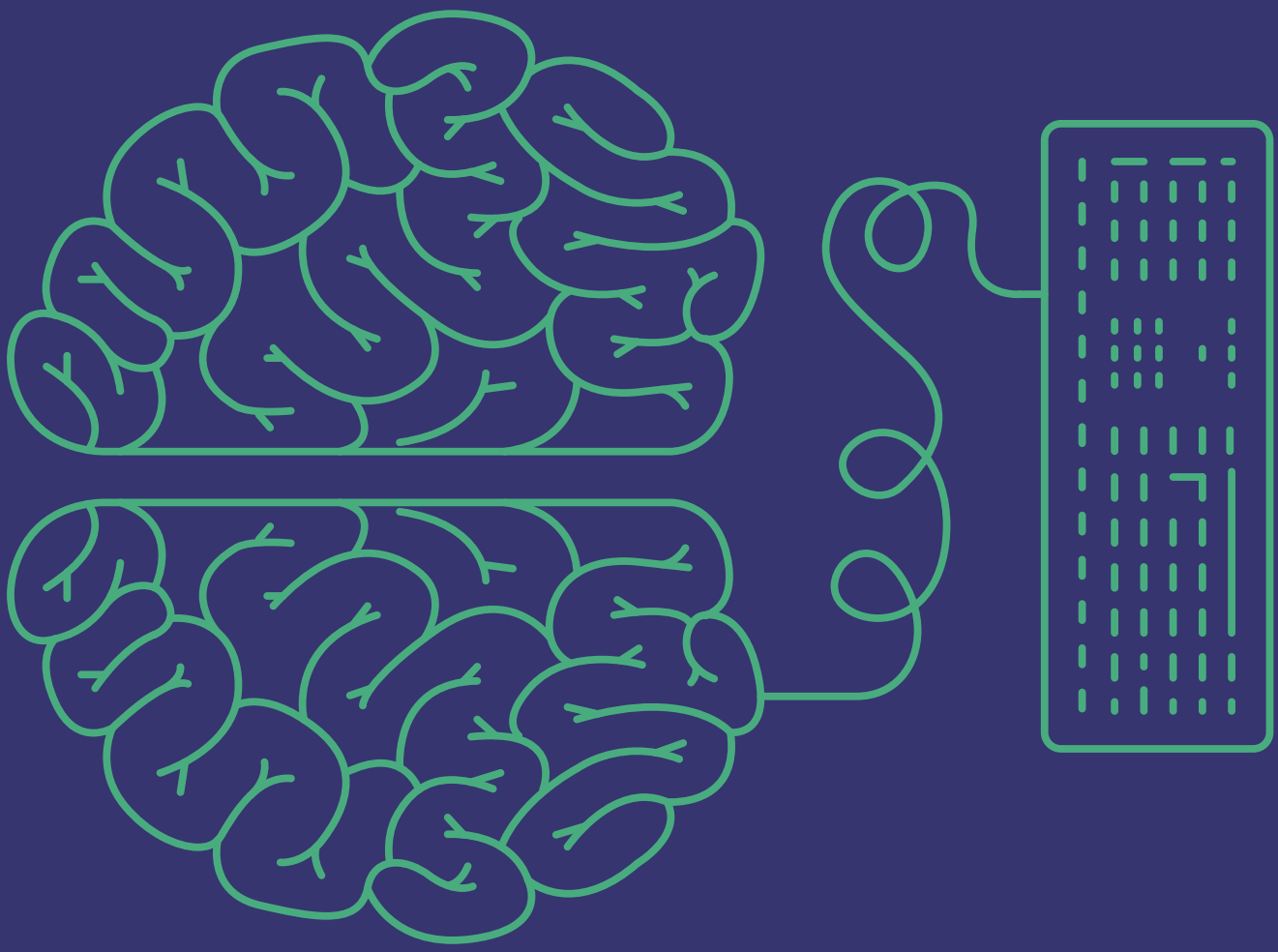
EFUTURES

The UK Landscape in Artificial Intelligence and Brain-Inspired Computing Hardware: the potential for establishing a new Centre of Excellence

August 2021

GIACOMO INDIVERI,
UNIVERSITY OF ZURICH AND ETH ZURICH

WALID NAJJAR,
UNIVERSITY OF CALIFORNIA, RIVERSIDE



EFUTURES

The UK Landscape in Artificial Intelligence and Brain-Inspired Computing Hardware: the potential for establishing a new Centre of Excellence

Table of Contents

Executive Summary

The world we live and work in is changing rapidly, with our daily lives more digitally enhanced and online-dependent than ever. This change has been made possible by the steady progress made in computing technologies. While appropriate new artificial intelligence (AI) algorithms have emerged, operating these algorithms is pushing current technology to its limit.

The current systems do not possess the capability that is imminently needed. The demands of AI neural networks, and of deep learning techniques that require thousands of petaflop-days to train, represent enormous challenges for computing technologies; as well as placing an unprecedented strain on energy consumption, and increasing carbon dioxide emissions.

The challenge facing us – of how to increase system capability to support the growing technological demands – has been the underlying agenda for a radical yet necessary paradigm shift in computer architecture, seeing it move towards a form of what we term '**brain-inspired computing**'. **This shift has been reflected in the creation of a number of major projects across the globe**: in Europe, the Human Brain Project; in the United States, with the BRAIN initiative; and in China, an ambitious AI development plan. The need for this radical paradigm shift in computing is urgent.

We must enable the sustainable and pervasive computing technologies required by society. Looking at international activities makes it clear that the countries leading on AI have taken the initiative in this domain, and it is this Report's strong recommendation that the UK should do the same.

Crucially, the activity on brain-inspired computing is central to the UK's recent published National AI Strategy (September 2021). The UK has an impressive commercial semiconductor expertise in AI-based computing, and the recommendations outlined in this report can directly contribute to the "UK's compute capacity needs to support AI innovation, commercialisation and deployment."

To enable an initiative on the necessary scale, this report recommends the establishment in the UK of a novel interdisciplinary centre of excellence for AI technologies and neuromorphic (or 'brain-inspired') computing.

The UK has some leading strengths in large-scale neural computing systems as well as in algorithmic development for AI-dedicated hardware – including material science, nanoscale memristive devices, analogue/digital circuits and systems design, and multi-core architectures. These strengths are the foundation upon which a Centre could be constructed.

This report has analysed in detail the UK landscape, reading peer-reviewed papers and interviewing UK researchers from seventeen institutions. As well as an analysis of representative scientific publications and (virtual) meetings, there have been tours of some institutions' facilities.

After reviewing the UK's outputs and abilities in this area, it is evident that for these national abilities to be fully harnessed to realise the potential in the UK to meet the considerable digital demands of twenty-first century, **a national programme is needed.**

This programme should coordinate research, nurture close and sustained communication across cogent disciplines, and develop efforts in this domain. **Such an endeavour would likely result in making the UK extremely competitive at the global level**.

The national programme would be instrumental in (a) promoting the growth of the research community; (b) in developing innovative and competitive AI technologies; and (c) in creating an industrial exploitation strategy to form synergies between the advanced research institutions and the many small, medium and large enterprises present in the UK that develop and use AI technologies. We include a series of recommendations within that, if addressed successfully, would ensure the success of such an enterprise.

Preface

We are delighted to present this independent report on the UK's current capabilities and future potential in energy efficient neuromorphic computing. The report emerged from detailed discussions between UKRI-EPSRC and the UK electronics systems community about the enormous potential of a radical new technology: brain-inspired, neuromorphic, computing, to transform the Artificial Intelligence (AI) landscape and spawn new industries, all while accelerating the adoption of AI – a government priority (and a central theme of the UK's new AI Strategy).

The community and EPSRC recognised early on that the UK has a leading position in this field, hosting some of the world's pre-eminent researchers, and a growing number of start-up companies laying the groundwork for a new industrial sector.We are delighted that the independent international experts who wrote this report concur. They describe compellingly the promise of this exciting new technology to increase massively the energy efficiency of computing. **They highlight where the UK holds a lead, and where it can leverage that to disrupt and innovate.**

They conclude that now is the time for the UK to capitalise on its strengths and create a Centre of Excellence as a focus and stimulus – bringing together the somewhat disparate groups working in the field; enabling the best science; training the next generation of researchers and innovators and, importantly, accelerating innovation and translation to industry.

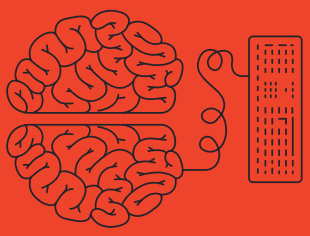
What is proposed is distinct from, but highly complementary to, existing initiatives such as the Turing Institute. The new entity would focus on wholly new ways of computing – ways that take inspiration from the brain rather than being further developments of existing digital technologies – and would initially focus on hardware: new types of computer that are hugely more power efficient than existing digital systems.

While AI would be the principal beneficiary, the new technology would disrupt other microelectronics sectors, including sensors, satellites, automotive, and industrial process control – all important UK industrial sectors.

Crucially, while current AI technologies consume ever more power (our best digital systems are at least a factor of a million times more power-hungry than the human brain), neuromorphic systems have a crucial role to play in delivering the UK's Net Zero agenda. Now is indeed the time.

The report was co-ordinated by the UKRI-funded eFutures2.0 network (https://efutures2.com/), which is a collaboration of various UK universities (Bristol, Cambridge, Edinburgh, Glasgow, Imperial College London, King's College London, Liverpool, Manchester, Newcastle, Queen's University Belfast, Sheffield, Southampton, Sussex, York) and companies (ARM, Huawei and PragmatIC).

It acts to highlight and promote the benefits of electronic systems, explore emerging technology challenges and create community events. To date, it has organised a series of activities in AI for vision systems, healthcare for low income countries, cybersecurity, future computing technology and spintronics.



'Neuromorphic systems have a crucial role to play in delivering the UK's Net Zero agenda'

Professor Tony Kenyon,
University College London

Professor Roger Woods,
Queen's University Belfast

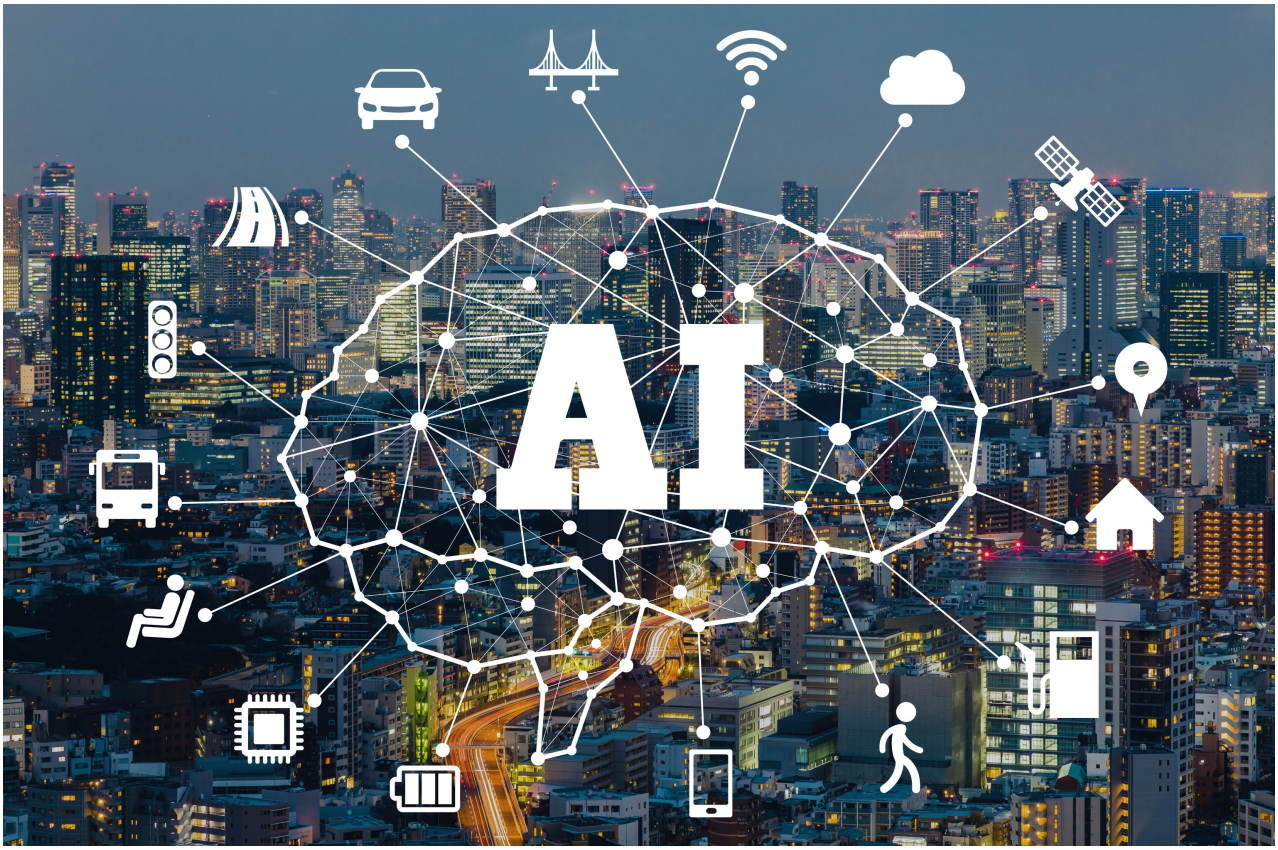
Opportunities and challenges for brain-inspired AI computing

The Age of AI

A technological revolution is in the making. It encompasses almost all aspects of our society, from education to health, from finance to automation, from transportation to climate change. Computing technologies have become pervasive and, as a result, increasingly large amounts of data are being produced every year. New, cutting edge Artificial Intelligence (AI) algorithms and data-science methods are needed to both exploit the opportunities, and to cope with the demands that are emerging with this revolution. AI algorithms, which typically employ neural network deep learning techniques to solve pattern recognition have been extremely successful in extracting information from large amounts of data **[1].** The methods used, however, to develop the latest and most powerful networks, such as GPT-3 **[2]** require thousands of petaflop-days to train (over 10^{23} floating-point operations). It has been estimated that the multiple training sessions used to develop GPT-3 required "9,998 total days" worth of GPU time (more than 27 GPU-years). Taking all these runs into account, the researchers estimated that building this model generated over 35 tonnes of carbon dioxide emissions: more than the average American adult will produce in two years."**[3]**

The size, the computational power, and the energy requirements of these networks are steadily increasing. But clearly, this trend is not sustainable [4].

[1] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015); Schmidhuber, J. Deep Learning in Neural Networks: An Overview. Neural Networks 61, 85–117 (Jan. 2015).
[2] Floridi, L. & Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 1–14 (2020).
[3] From the Forbes article "Deep Learning's Carbon Emissions Problem", https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/3
[4] Big Data Needs a Hardware Revolution. Nature 554, 145–146 (2018).



The Age of AI

The need for a radical paradigm shift

The strategies currently used to design, run, and train these networks use conventional computing technologies based on the classical "von Neumann architecture".

This scheme comprises separate memory and processing units which transfer data across a shared serial bus as quickly as possible.

On the other hand, human brains, which clearly outperform AI computing systems in terms of the amount of training data, power consumption, and adaptability to novel unexpected conditions, use a radically different computing paradigm.

In animal brains, memory and computation are co-localized in massively parallel arrangements of slow and unreliable computing elements. Current AI and deep learning methods are inspired by the architecture of the cerebral cortex. Although there has been tremendous progress in neuroscience, in terms of better understanding the principles of computation used by the brain, there is a massive gap between relating natural intelligence to AI computing systems, both in terms of the physical computing substrate and the organizing principles.

Current AI paradigms, and the supporting computing technologies, cannot sustain the demands that are coming from this revolution. New research programs aimed at creating novel computing hardware and methods that can go well beyond the current AI solutions have recently begun in several international labs as well as national ones. Investing in research and development in this area has very high potential for future technology transfer and exploitation, which can reap the benefits of a growing global market in edge computing and the Internet of Things.

The global landscape in brain-inspired computing

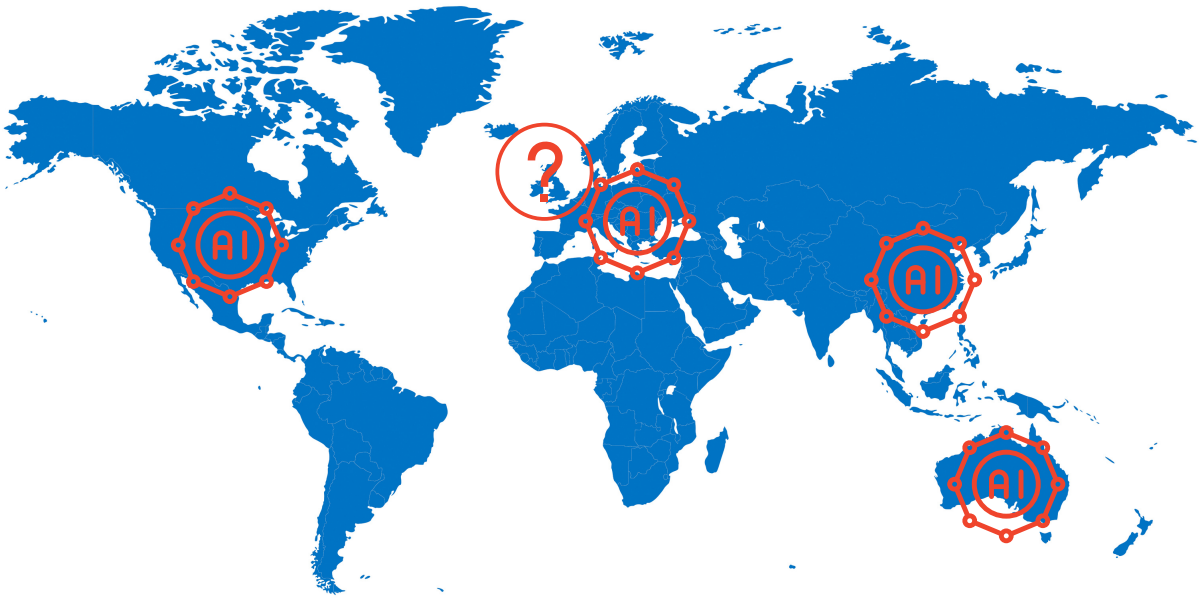
Early remarkable initiatives that recognized that studying how brains work could address the great computing challenges of our age started in Europe, with the Human Brain Project (HBP); in the United States, with the BRAIN initiative; and in China with an ambitious AI development plan. The EU Human Brain Project is a large ten-year "flagship project" funded by the European Union, that started in 2013 with the goal of building methods, tools and the infrastructure to advance neuroscience, medicine, and computing. The brain-inspired computing aspect of the project was developed within the neuromorphic computing pillar.

There were two major platforms being developed: the BrainScaleS-1 wafer scale system and the SpiNNaker system **[5].** Notably, the SpiNNaker system was developed and is currently still being maintained and further developed by the group of Prof. S. Furber, at the University of Manchester, UK **[6]**.

The SpiNNaker system currently represents one of the most complex and computationally powerful brain-inspired computing architectures. It was designed to support the simulation of large-scale spiking neural networks and brain models, and is serving this purpose remarkably well. It provides researchers with a tool for developing computational models that can potentially be implemented with future emerging brain-inspired computing technologies.

US BRAIN is a joint public-private research initiative announced by the Obama administration in 2013, with the goal of supporting the development and application of innovative technologies that can create a dynamic understanding of brain function. While this initiative supports the development of novel electronic technologies, the focus is mainly on creating methods and tools for supporting experimental neuroscience and basic research. In this respect, this initiative is not directly targeting work on AI and computing architectures.

[5] Furber, S., Galluppi, F., Temple, S. & Plana, L. The SpiNNaker Project. Proceedings of the IEEE 102, 652–665. issn: 0018-9219 (May 2014)
[6] SpiNNaker: A Spiking Neural Network Architecture (eds Furber, S. & Bogdan, P.) isbn: 978-1-68083- 653-0 (Boston-Delft: Now Publishers, 2020)



“There is opportunity for the UK to excel in this field”

China's investments in AI, on the other hand, are directed toward market-driven computing and applications. In 2017, the State Council of China released the "Next Generation Artificial Intelligence Development Plan", which outlines China's strategy to build a domestic AI industry worth nearly US$150 billion in the next few years and to become the leading AI power by 2030. This represents probably the largest national investment plan in research and development of AI that encompasses multiple levels, ranging from research in materials science and devices to the development of high-performance computing systems. It targets applications that range from large data-servers for the digitization and processing of large volumes of information, to low-power intelligent edge-computing and brain-inspired sensory information processing systems.

Since these initial efforts, many other similar national and international initiatives have been launched (e.g., in France, Germany, Canada, Australia, and many other Asian countries). All these early initiatives, however, fuelled the development of the "classical" AI methods, which have been using conventional computing architectures based on Central Processing Units (CPUs) and Graphical Processing Units (GPUs), and have focused mainly on incremental improvements of standard neural network models and academic benchmarks of no practical importance for real-world applications, such as the MNIST data-set of handwritten digits.

Challenges with conventional AI approaches

Given the mismatch between the serial time-multiplexed nature of conventional (so called "von Neumann") computing architectures and the massively parallel nature of neural networks, the vast majority of the AI algorithms and their supporting hardware developed so far still face many open challenges. The most critical of these challenges lies in the amount of energy required to run these very large networks; but other challenges need to be resolved urgently, for example the requirement of very large labeled annotated data-sets, the lack of "explainability" of both model parameters and outputs, and the potential lack of data-security.

The current trend to improve the performance and accuracy of these networks is that of increasing the size of the model and the number of parameters they use. This approach does not address the challenges mentioned above. On the contrary, it increases the amount of petaflop-days (and corresponding carbon footprint) used for training these algorithms, so it represents a short-term solution that will not scale.

The UK is very well positioned to address these challenges, by supporting the existing community of world-leading experts in low-power, brain-inspired hardware computing technologies for AI.

Investments in this area, and the creation of an initiative to coordinate the broad set of expertise present in UK has large potential to generate significant public benefit.

The Report Process

The UK has already a very comprehensive set of labs and institutions that are doing state-of-the-art research and development within the field of hardware for AI. To assess the strengths and weaknesses of the UK landscape in these areas of research, we were invited as independent international experts and asked to study the work of a sample of representative UK researchers and research institutions, by reading their scientific publications, holding (virtual) meetings, and arranging (virtual) tours in some of their labs.

There were 17 institutions in total that were involved:

- Imperial College London
- King's College London
- Liverpool John Moores
- Loughborough University
- Middlesex University
- Queen's University Belfast
- University College London
- University of Edinburgh
- University of Exeter
- University of Hertfordshire
- University of Hull
- University of Manchester
- University of Newcastle
- University of Sheffield
- University of Southampton
- University of Sussex
- University of Ulster

Table 1 lists the meeting dates held with researchers from some of these Universities, and the colleagues met.

It quickly became clear that there is a very broad spectrum of areas of expertise and themes covered by these researchers.

The research and development activities in this area span topics that range from material science and nanoscale memory device development to analog/digital circuits and systems design, to multi-core power and architecture optimization for very large-scale neural computing systems and software development for AI-dedicated hardware, all the way to application development and deployment for Internet of Things use cases.

Table 1: Meetings with UK University Research Teams

University	Meeting date	Participants
Southampton	2 Dec. 2020	Themis Prodromakis, Geoff Merrett
Imperial College London	3 Dec. 2020	Christos Bouganis, George Constantinides, Wayne Luk
King's College London	3 Dec. 2020	Bipin Rajendran, Osvaldo Simeone
University College London	3 Dec. 2020	Tony Kenyon, Adnan Mehonic, Alex Shluger
Edinburgh	8 Dec. 2020	Alister Hamilton, Srinjoy Mitra
Manchester	9 Dec. 2020	Simon Davidson, Piotr Dudek, Steve Furber, Jim Garside, Oliver Rhodes, Jayawan Wijekoon
Queen Belfast	11 Dec. 2020	Dimitrios Nikolopoulos, Hans Vandierendonck, Roger Woods

Table 2 highlights some of these areas, and lists a subset of the UK University and research labs that have strong expertise in them:

Table 2: Areas of expertise covered by the UK University Research Teams. The institutions listed are only a representative subset with very strong presence in the area of research outlined.

Area of research	Institutions
Materials and devices for AI memory technologies	London Center for Nanotechnology (University College London, Imperial College London, King's College London), University of Southampton
FPGA and processing micro-architectures optimization for AI	Queen's University Belfast, University of Southampton, Imperial College London
CNN sensors and processors	University of Manchester, Imperial College London
DNN hardware for learning and inference	Imperial College London, University of Southampton, King's College London
Brain-inspired neuromorphic computing hardware	University of Manchester, University of Edinburgh, University of Southampton
SNN computational modeling on standard computing platforms	University of Sussex, King's College London
Theoretical and computational neuroscience for AI	University College London (Gatsby)
AI demonstrators and application	University of Edinburgh, University of Manchester

FPGA: Field Programmable Gate Arrays; **CNN** Convolutional Neural Networks; **DNN**: Deep Neural Networks; **SNN**: Spiking Neural Networks

The opportunity for the UK to excel in this field

On one hand the breadth of skills and competences present in the field of brain-inspired and AI hardware is an advantage, as the UK has all the expertise required to develop a full computing stack, that ranges from materials and devices to applications and algorithms (e.g., see Fig. 1). On the other hand, creating a research programme that can coordinate the broad set of activities required to co-design brain-inspired AI hardware systems and applications at all these levels is going to be extremely challenging.

To best understand what these challenges entail it is useful to highlight the differences between the classical "stored program" computing stack and the "neuromorphic" computing stack.

The Stored Program Computing Stack

The tremendous growth of computing technology that was witnessed in the past half century is due, to a very large degree, to the layering of abstractions, referred to as the computing stack, allowed by the simplicity of the stored program computing model. These abstraction layers bridge the gap between electronic devices and end-user applications.

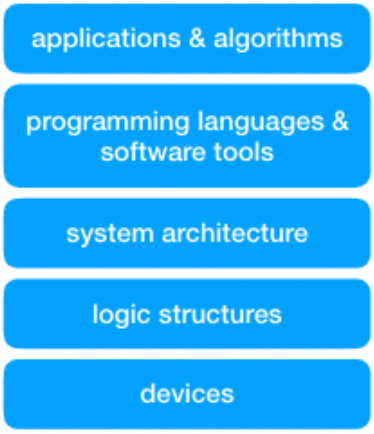


Figure 1: The stored program computing stack

Each level provides the primitives that are used by the level above it, hence the efficacy and performance of each level of the stack depends, to a very large degree, on the suitability of the primitives provided by the stack level below it.

One of the benefits of the computing stack is that it allowed the concurrent development and evolution of each layer while maintaining inter-operability and backward compatibility. This has also resulted in an accelerated technology transfer from research to commercial products.

The Neuromorphic Computing Stack

While neuromorphic computing is a radical departure from the traditional stored programme model (SPM), a similar layering of abstractions is to be expected by the industry and the market. It is important, therefore, that the neuromorphic computing model be viewed in the same framework to foster rapid adoption and inter-operability within its own computing stack.

As it did in the context of the SPM, we believe that a computing stack will develop organically within neuromorphic computing (see Fig. 2). The specifics of each individual layer of the stack will get defined and evolve over time. However, the end layers are akin to fixed-points: at the bottom a layer that includes the devices supporting the computation, at the top level a layer supporting the end-users applications.The intermediate layers must bridge the gap between the applications and the devices.

An important caveat, however, is that while the SPM is an ultra general purpose model (i.e Turing equivalent) that supports the widest possible range of applications, the neuromorphic computing structures target a much narrower range of algorithms. In fact, the neuromorphic computing structures maybe described as programmable application domain specific computing structures; and hence may be more similar to Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICS).

This narrower range of application styles allows for, and nearly requires, a tighter integration between the layers of the computing stack. It is therefore very important that application developers and end users be involved from the inception of the research effort.

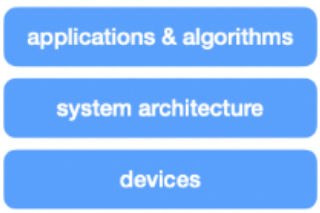


Figure 2: The neuromorphic computing stack

A national programme on emerging technologies for AI

Given the urgent need for a radical paradigm shift in computing to obtain sustainable and pervasive computing technologies, the international attempts currently being made in this domain, and the strong base of world-leading experts distributed across multiple UK Universities and research institutions, **it is an ideal time to capitalize on the UK assets and start a national education, research and development program on brain-inspired AI technologies and neuromorphic computing**.

The time is ripe for coordinating the UK community working on these themes and fostering its growth, to lead the development of innovative and competitive AI technologies.

A national coordinated effort in this area would be instrumental for consolidating the existing know-how, for exploiting the advancements already made in this field by the UK institutions, for supporting technology-transfer, and for creating a successful industrial exploitation strategy able to form synergies between the advanced research institutions and the many small, medium and large enterprises (e.g., such as DeepMind, ARM, Graphcore, Intrinsic, ARC instruments, Sonet, etc.), for eventually generating revenue from future advanced AI computing systems and products.

Relevant areas to involve

While there is no doubt that AI technologies will have tremendous impact on a most, if not all, areas of human activity, this impact is, as of yet, neither measured nor quantified. The transition from science to engineering is known to be challenging, tedious and time-consuming. Similarly, the transition from engineering to applications can take similar paths. It is critical, therefore, that applications stake holders be "in the loop" in the selection and formulation of target applications.

Performance

Obviously, the performance of the system is of primary importance. This includes the training time, pre-deployment, as well as the response time post-deployment.

Energy Consumption

This has been the main obstacle to the growth in size of traditional AI and machine learning [ML] systems.
Novel technologies should target, at least, an order of magnitude reduction in total energy consumed per training task. One avenue for achieving this goal may be migrating tasks to edge or mid-level nodes.

Security and Privacy

Security and privacy **[7]** have become primary objectives for all information technology systems and more so for AI/ML systems. As the IT industry has been discovering, retrofitting security and privacy criteria into existing systems is, when possible, very tedious, costly, and prone to more errors and vulnerabilities. These objectives should be included and addressed at every stage of the system design.

Explainable AI (XAI)

Explainable AI [XAI] reverses the black box concept of traditional AI, making the results of the"algorithm" understandable by humans. The relevance of XAI goes beyond the legal and ethical requirements, it can improve the user experience of a product or service **[8]**. It is expected that hardware support for XAI, at both the training and deployment phases, to reduce the overhead costs and accelerate the explanation **[9].**

[7] Within this rubric we are including all aspects of algorithmic social bias (ethnic, racial, religious, sexual orientation, etc.

[8] https://www.darpa.mil/program/explainable-artificial-intelligence

[9] An excellent taxonomy & summary of challenges in XAI can be found in: Barredo Arrieta, A. et al.

Real World Inputs and Outputs [I/O]

The hardware design, and evaluation, of novel computing platforms often fails to account for how the data from outside the system, on storage or from sensors, would be marshalled and made available to the system.

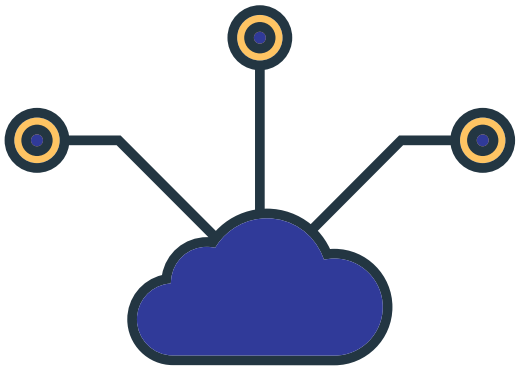
Novel hardware technologies must work efficiently with current memory and storage technologies, and be able to carry out on-line data processing, streaming directly from the sensors and using compatible communication and data-representation protocols.

Cloud and Edge Computing

So far, most traditional AI/ML systems have been designed for massively large power hungry cloud-based systems. Examples include Google Cloud TPU v3 **[10]** and Intel's recent acquisition of Habana Labs, maker of Gaudi (deployed in AWS) and Goya **[11]**.

Recently, Google Research introduced the Google Coral **[12]** designed for integration with edge-based sensing and computing nodes and based on the Google Edge TPU **[13]**.

It is expected that the emphasis on accelerators for AI computing at the edge will keep on growing. The main advantages and objectives are to (a) reduce the load on the cloud servers by distributing the tasks; (b) reduce the reliance on high bandwidth to transfer raw data; and (c) increase the levels of security and privacy by transmitting processed rather than raw data to the cloud servers. For some applications, mid-level computing nodes maybe be able to achieve the same objectives as edge computing nodes.



10. https://cloud.google.com/tpu/docs/system-architecture
11. https://habana.ai/
12. https://www.coral.ai/
13. https://cloud.google.com/edge-tpu/8

Neuromorphic behaving agents

Brain-inspired computing technologies must be able to exhibit intelligence by producing adequate behavioural responses, based on the external stimuli and internal state of the computing system.

When interaction with the environment comes into play, the computing system should become an autonomous agent.

An autonomous agent - in this sense - is a miniaturised, intelligent, embedded processing system that can sense signals from, and interact with, the environment to solve real-world problems; without requiring access to remote or high-energy computing resources (such as standard computers or data-centres).

Existing Analogous Initiatives

In January 2021, the Semiconductor Research Corporation (SRC) released "The Decadal Plan for Semiconductors – A Pivotal Roadmap Outlining Research Priorities."**[14]** Of their five priorities, the first is "The Analog Data Deluge - Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive, and reason".

The annual investment need is US$600 million throughout this decade to pursue analog-to-information compression/reduction, with a practical compression/reduction ratio of $10^5:1$ for practical use of information more analogous to the human brain.

The recently released final report of the National Security Commission on Artificial Intelligence (NSCAI, 2021) **[15]** stresses the urgency of extensive government investments in all aspects of AI R&D.

Even a summary of this extensive report is beyond the scope of this document, however, notable highlights include:
(1) The establishment of a National Technology Foundation [NTF] as a sister organisation to the National Science Foundation [NSF] with a mandate to accelerate the transition from science to engineering by focussing on eight priority areas, including AI, Semiconductors, and Advanced Hardware.
(2) An immediate investment of US$30 billion in microelectronics R&D and manufacturing.
(3) Double annual non-defence AI R&D funding to reach US$32 billion by 2026.

[14] https://www.src.org/about/decadal-plan/
[15] Available at https://www.nscai.gov/2021-final-report/

Intel's Loihi is a 128-core self-learning neuromorphic research chip (2017, 14nm technology), whose architecture is optimized for SNN algorithms.

Poihoiki Springs is a mesh of 768 Loihi chips that implements a 100 million spiking neurons.

The Human Brain Project [HBP] is building a research infrastructure to help advance neuroscience, medicine, and computing. **[16]**

The EBRAINS Research Infrastructure is part of HBP including SpiNNaker (Manchester, UK) and BrainScaleS (Heidelberg, Germany).

In 2017, China set national goals to lead the world, by 2030, in a number of technologies including AI and microelectronics.

The parameters and challenges of the competition between the US, and the West in general, and China in advanced technologies are extensively discussed in the NSCAI 2021 report. **[17]**

[16] https://www.humanbrainproject.eu/en/silicon-brains/
[17] Available at https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

Seven Specific Recommendations

1 Mission Statement

The foremost recommendation is to define a global mission statement with a vision that can unite all interested parties to work together toward a common goal. The mission statement should be broad enough to be able to include all experts in the UK working in the field of brain-inspired hardware for AI, but specific enough to focus the research and development activities to those that are strictly necessary for accomplishing the centre's goals.

2 Broaden involvement to progress the stack

While it is clear that UK has a very strong presence at the base of the neuromorphic computing stack of Fig. 2, namely on materials, devices and circuits for AI, the middle part of the neuromorphic computing stack has a smaller set of representatives, albeit highly expert and influential.

Importantly, the top part of the stack needs to be further developed to be able to pursue a holistic co-design approach. To increase the strength of the intermediate levels of the neuromorphic computing stack, and to exploit the outcome of the material science and memory device developments, it will be useful to ramp up the UK's activities and education in application specific integrated circuit (ASIC) design.

To exploit the results from the bottom layers of the neuromorphic computing stack and enable the development of AI applications and products it will be useful to invite more computer scientists, software AI architects, and computational modelling experts interested in applying their know-how to the centre's new vision.

3 Close co-operation between hardware & software design

The national program should support research projects that involve and bring together two or more layers of the neuromorphic stack, involving multiple centres of research strength in the UK, in each of the computing stack levels.

The history of computing is littered with very sophisticated parallel architecture designs that (a) did not consider how the software layer would interface to the architecture and/or (b) focused on a very small set of trivial kernels (matrix multiplication being the all time favourite) to demonstrate the performance of the proposed machine(s).

A well designed hardware software interface will provide the users with multiple paths to program the machine (compilers) as well as supporting an effective management of the hardware resources (operating system) that allows fair allocation of resources to multiple tasks, whether within the same application or across applications.

4 Draw in Neuroscience

As this program will require the creation of a radically different and novel computing paradigm based on brain-inspired computing principles, it should exploit the large number of world-leading experts present in the UK in theoretical, computational, and experimental neuroscience.

Developing a deep understanding of the computational principles of the brain will be crucial for understanding how to carry out low-power and robust computation using massively parallel arrays of inhomogeneous and low-resolution electronic components, similar to the way the brain uses synapses and neurons.

5 Foster meaningful & sustained multi-disciplinarity

Creating a new brain-inspired computing technology with a novel unconventional computing stack is something that will require interdisciplinary work and interactions among experts in different domains.

It will be important to set up an infrastructure that will encourage cross-domain fertilization, educational opportunities for learning to understand each other's field of expertise (e.g., to understand and "speak the same language"), and importantly to train a new generation of young investigators to be competent in all relevant topics from multiple and diverse disciplines, crossing traditional boundaries between standard subjects, and gridlocked mindsets.

6 Identify Key Objectives

In terms of areas to involve, the program should make Security, Privacy, XAI, and Neuromorphic Behaving Systems primary objectives in the conception and execution of novel neuromorphic computing projects.

In addition it should provide incentives to the research teams to include some variety of active application domains stake-holders.

7 Define your measures

To properly assess the advancements made in research and development, it is crucial to define from the onset of the project quantifiable goals and realistic milestones.

These will need to take into account fundamental research aspects, the industrial strategy goals, and exploitation and innovation aspects. It will be important to define the nature of the IP that will be generated and ways to quantify its effect on society.

A UK Centre of Excellence on brain-inspired hardware for AI

In our opinion, the best way to succeed in this effort, and implement our recommendations, is to create a UK Centre of Excellence on brain-inspired hardware for AI.

By establishing such a centre and giving it a clear mission statement it will be possible to bring all interested parties together and coordinate their activities to reach the program's objectives in an efficient way.

The centre will be responsible for demonstrating the results of the research funded, and act as the link between the researchers and the UK stakeholders.

Photo: Anatolii Nesterov on Unsplash

Expertise of the Authors

Giacomo Indiveri is a dual Professor at the Faculty of Science of the University of Zurich and at Department of Information Technology and Electrical Engineering of ETH Zurich, Switzerland. He is the director of the Institute of Neuroinformatics of the University of Zurich and ETH Zurich. He obtained an M.Sc. degree in electrical engineering in 1992 and a Ph.D. degree in Computer Science from the University of Genoa, Italy in 2004. An Engineer by training, Indiveri has also expertise in neuroscience, computer science, and machine learning. He has been combining these disciplines by studying natural and artificial intelligence in neural processing systems and in neuromorphic cognitive agents. His latest research interests lie in the study of spike-based learning mechanisms and recurrent networks of biologically plausible neurons, and in their integration in real-time closed-loop sensory-motor systems designed using analog/digital circuits and emerging memory technologies. His group uses these neuromorphic circuits to validate brain inspired computational paradigms in real-world scenarios, and to develop a new generation of fault-tolerant event-based neuromorphic computing technologies. Indiveri is senior member of the IEEE society, and a recipient of the 2021 IEEE Biomedical Circuits and Systems Best Paper Award. He is also an ERC fellow, and recipient of three European Research Council grants.

Walid A. Najjar is a Professor and Chair of the Department of Computer Science and Engineering at the University of California Riverside. He received a B.E. in Electrical Engineering from the American University of Beirut (1979), and M.S. (1985) and Ph.D. (1988) in Computer Engineering from the University of Southern California. From 1989 to 2000 he was on the faculty of the Department of Computer Science at Colorado State University, before that he was with the USC-Information Sciences Institute. He is an ACM Distinguished Scientist and Fellow of the IEEE and the AAAS. His areas of research include computer architectures and compilers for parallel and high-performance computing, embedded systems, FPGA-based code acceleration and reconfigurable computing. In recent years, he has worked extensively on the design and implementation of FPGA-based high-performance code accelerators for a wide range of applications such as computer vision, data mining, bioinformatics, databases, and data analytics.

Bibliography

1. LeCun, Y., Bengio, Y. & Hinton, G. "Deep learning", Nature 521,436–444 (2015).

2. Floridi, L. & Chiriatti, M. "GPT-3: Its nature, scope, limits, and consequences", Minds and Machines, 1–14 (2020).

3. Schmidhuber, J. "Deep Learning in Neural Networks: An Overview", Neural Networks 61, 85–117 (Jan.2015).

4. "Big Data Needs a Hardware Revolution", Nature 554,145–146 (2018).

5. Furber, S., Galluppi, F., Temple, S. & Plana, L. "The SpiNNaker Project", Proceedings of the IEEE 102, 652–665.issn: 0018-9219 (May 2014).

6. SpiNNaker: A Spiking Neural Network Architecture(eds Furber, S. & Bogdan, P.) isbn: 978-1-68083-653-0 (Boston-Delft: now publishers, 2020)

7. Barredo Arrieta, A.et al."Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunitiesand challenges toward responsible AI", Information Fusion 58, 82–115.issn: 1566-2535. https://www.sciencedirect.com/science/article/pii/S1566253519308103 (2020)

Contact

web. www.efutures2.com
email. beth.mcevoy@qub.ac.uk
twitter. @efuturesuk

EFUTURES

eFutures is funded by the UK Engineering & Physical Sciences Research Council

UKRI Engineering and Physical Sciences Research Council