Findings and Recommendations of the May 2022 US-UK AI Workshop

May 3-4, 2022

Jeannette Wing (Columbia University) and

Michael Wooldridge (University of Oxford & Alan Turing Institute)

Sponsored by the
National Science Foundation (NSF) and
Engineering and Physical Sciences Research Council (EPSRC)

Findings and recommendations from a two-day virtual summit funded by the National Science Foundation (NSF) and the Engineering and Physical Sciences Research Council (EPSRC) to enhance collaborative efforts on research and innovation between the US and UK in developing AI solutions for mutual benefit.

Executive Summary

On September 25th, 2020, the United States and United Kingdom signed a joint declaration on cooperation in Artificial Intelligence (AI), signaling an intention to work together for mutual benefit on this rapidly evolving and strategically important area of research and development. In furtherance of this declaration, the NSF in the US, and its counterpart the EPSRC in the UK convened a joint virtual workshop on AI in May 2022. This workshop brought together some 50 leading researchers and practitioners in AI from both nations, with the goals of sharing knowledge and experience, and identifying and prioritizing possible areas for US-UK cooperation going forward. This report summarizes the findings and recommendations of that workshop.

While the prior expectation was that participants would focus largely on specific research areas, during the event participants also devoted considerable time to discussing the wider societal issues raised by present and likely future developments in AI. The report thus identifies a range of technical research topics, but in addition makes recommendations surrounding issues such as the societal impact of AI, and how systemic problems such as the lack of diversity in AI might be addressed.

In terms of specific research topics, participants were vocal in their view that the now well-known issues of trustworthy AI and explainable AI remain very prominent, and deserving of further research. Trustworthiness touches on issues from human-AI interaction to the formal verification of AI systems, while explainability relates to the issue that contemporary machine learning techniques result in "black box" systems, from which it is difficult or impossible to distill any human-comprehensible rationale for a system's behavior. Both of these issues represent substantial impediments to the wider takeup of AI, particularly in safety-critical applications. Participants noted that many prominent recent AI systems (such as OpenAI's GPT-3) appear to have impressive capabilities in some problem domains, and yet we have difficulty in understanding exactly what the limits of these capabilities are. Work is therefore required on systematically benchmarking the capabilities of such systems. Much discussion in the workshop focussed on the issue of computational resources and data used for machine learning systems, and in particular, the fact that state-of-the-art machine learning systems require extraordinarily large data sets to train models on large computational infrastructure; research is thus required to enable machine learning with small data sets. While machine learning approaches are currently predominant in AI research, participants noted that

combining such approaches with other AI technologies (e.g., probabilistic reasoning, symbolic reasoning) is a natural direction for investigation. Finally, new hardware paradigms such as quantum computing offer the potential to transform the capabilities of AI systems.

Turning to the societal impacts of AI, participants strongly voiced their desire to pursue beneficial applications of AI research, and to support this, advocated the dissemination of benchmark problems and data sets in areas such as climate, health, transport, agriculture, food production, and the urban environment. Participants also identified the importance of the human-centered AI paradigm, which explicitly considers the role that humans play in all aspects of the AI ecosystem. Education was also seen as a key issue: both the need to educate the public about AI, as well as the need to disseminate technical information about AI to researchers.

In terms of specific application areas, finance, public health, the natural sciences, climate change/sustainability, and assistive living were all identified as natural targets for application-oriented research.

Finally, participants strongly signaled the desire to see more done around the issue of diversity in AI. This meant not just addressing the well-documented lack of diversity within the AI research community, but tackling such issues as ensuring diversity in data sets and considering diversity in the design of AI systems. There was considerable discussion on the topic of diversity in the AI talent pipeline, and a range of measures were proposed to support this.

Taken together, the findings and recommendations from the workshop that are detailed in the remainder of the report represent a compelling snapshot of the predominant issues in AI as of mid-2022. It is particularly striking, and we believe reassuring, that workshop participants were just as concerned with the wider societal implications of AI – and of ensuring that AI achieves globally beneficial outcomes – as they were with the bread-and-butter technical issues that we might have expected to dominate discussions.

Context

Artificial Intelligence (AI) is both a scientific and an engineering discipline with origins tracing back to the mid-20th century. AI is generally concerned with building artifacts that act on the behalf of humans in situations which require a certain degree of sophisticated intelligence. Since the genesis of AI research in the 1950s, a range of techniques proposed for AI succeeded to a greater or lesser extent in different domains. Symbolic AI – the dominant AI paradigm until the late 1980s – focused around the idea that intelligence is primarily a problem of knowledge, and the way to construct an artifact capable of intelligent action in some domain is to equip that artifact with the relevant knowledge about that domain. Although this approach demonstrated some notable successes, Symbolic AI largely failed with a range of tasks, particularly those involving perception and learning. Since then, Machine Learning (ML) established itself as the dominant paradigm for approaching AI solutions.

With ML, rather than explicitly telling a machine how to carry out some task, we show it what we want it to do and the machine learns how to carry out the task. So-called 'neural networks' are the key technology underpinning contemporary ML. A very old idea, neural networks have demonstrated remarkable successes this century. These successes are driven by new scientific developments ("deep learning"), the availability of large, high quality curated data sets, and the availability of increasingly inexpensive, yet powerful computer resources for training ML systems. Neural network technology scored a range of very high profile successes over the past decade: machines that learned how to beat human grandmasters in the game of Go and chess[1,2], systems exhibiting an unprecedented level of performance in the fundamental protein folding problem[3][ii], and programs automatically generating astonishingly lifelike images from textual prompts[4] are all wonderful examples of landmark developments.

Recent progress in AI is accompanied by enormous investments, with big tech companies in particular spending unprecedented sums of money to secure leading AI/ML researchers and to establish labs performing fundamental AI research. At the same time, the scale of contemporary AI systems – the data and compute resources required to build them – grows sharply, to the point where some contemporary ML models have required millions of dollars in

---

[1] https://www.deepmind.com/research/highlighted-research/alphago

[2] https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go

[3] https://alphafold.ebi.ac.uk

[4] https://imagen.research.google

compute resources to train them. As a consequence, there is a notable shift in the center of gravity of AI research from the public realm (universities, governmental organizations, and national research institutes) into the private realm (big tech in particular). The prospect of this key technology development outside of the public realm raises many legitimate issues of ethical and pragmatic concern for governments, universities, and funding agencies.

Recognizing the potential of this emerging new technology and the unique landscape in which fundamental AI research is now carried out, many national governments are responding with new initiatives. These initiatives aim to realize the manifest benefits of the new technology while mitigating its potential harms. In the United States, the National AI Initiative Act of 2020 became law on January 1, 2021[5], with the goal of providing a coordinated set of measures to support AI development and enable its beneficial applications. Similarly, the United Kingdom published a National AI Strategy in September 2021.

On September 25th, 2020, the United States and the United Kingdom made a declaration on "Cooperation in Artificial Intelligence Research and Development."[6] This compact is a recognition of the enormous potential benefits that AI enables for future economic growth, health and wellbeing, and national security. Both nations recognize that accomplishing such goals requires massive effort across many interdisciplinary fields of research to develop AI that is robust, accurate, adaptive, fair, and practicable. Furthermore, the ideation, cooperation, and implementation of AI in a productive manner for both nations requires an equally robust multilateral plan with continuous engagement.

In the furtherance of this declaration, the National Science Foundation (NSF) in the United States, and its UK counterpart the Engineering and Physical Sciences Research Council (EPSRC), organized a virtual workshop over two days in May 2022, involving the leaders of AI research in the US and UK. Participants were drawn predominantly from academia, but given the predominance of AI research in industry, a number of key industrial participants were also invited. The workshop aimed to identify priorities for possible US-UK collaborative research; to set an agenda to build upon both pre-existing US-UK collaborations as well as potential new research areas that benefit both nations; to establish practices for continuous engagement for

[5] https://www.ai.gov

[6] https://www.gov.uk/government/publications/declaration-of-the-united-states-of-america-and-the-united-kingdom-of-great-britain-and-northern-ireland-on-cooperation-in-ai-research-and-development/declaration-of-the-united-states-of-america-and-the-united-kingdom-of-great-britain-and-northern-ireland-on-cooperation-in-artificial-intelligence-re

progress updates, alterations to major plans, and sharing of new research and development breakthroughs; and to overall generally enhance cross nation communication within this domain.

This report summarizes the findings and recommendations arising from this workshop. The workshop not only set the framework for a coherent agenda on AI development and its application to the US and UK industries, but it also facilitated a structure for continued collaboration and communication between academic, industrial, and government domains. By convening over 40 AI researchers and leaders in one forum, key areas of research were identified for funding with a follow on agenda set to accomplish certain goals; the potential for advancement in knowledge in these fields of research is significantly improved.

U.S. National AI Initiative

Prioritize AI R&D
Grow and sustain U.S. research leadership and capacity

Strengthen AI Research Infrastructure
Enhance access to high quality data, models, and computing resources

Advance Trustworthy AI
Modernize governance and technical standards for AI-powered technologies, protecting privacy, civil rights, civil liberties, and other democratic values

Leverage AI for Government and National Security
Apply AI to improve provision of government services and national security

Promote International AI Engagement
Engage with like-minded allies to promote a global AI environment supportive of democratic values

Train AI-Ready Workforce
Provide AI-ready education at all levels: K-12, college, re-training, re-skilling, R&D workforce.

U.K. National AI Strategy

Investing in the Long Term Needs of the AI Ecosystem
Supporting proven growth levers:
Skills and Talent – R&D – Data – Compute – Finance and VC

Ensuring AI Benefits All Sectors and Regions
- Making AI accessible to industries and businesses across the whole country
- Using AI for the public benefit in service of bold missions

Governing AI Effectively
- Working with our international partners to embed AI governance values shared by the UK
- A world-leading domestic regulatory model that balances innovation and risk

The Workshop: Guiding Principles, Key Themes, and Structure

In order to organize the discussion and potential desired outcomes, we first established a few guiding principles and key questions as follows:

- **Strategic Prioritie**s: What are the key areas of research within the domain of AI? How can these areas of research be used to practically benefit both nations in the realms of defense, economic productivity, and citizen wellbeing? Which of these issues is most pressing and most readily addressable with current AI capabilities? Which projects are directly addressing these problems? What new projects need to be formed to tackle some of these issues? In which areas of AI research would there be differences in approach, focus, or implementation for the US and UK? Similarities?
- **Collaboration**: What are the current pain points in collaborative research between countries? What bottlenecks or opportunities exist across the industry and academic interface? How can we better facilitate cooperation and information flow across industries and nations? What can we learn from specific fields of research that have good collaboration across nations and academic realms? How would those apply to the individual fields of AI research or in total?
- **Agenda & Implementation**: Which projects can be immediately funded to address the identified pressing issues? What are realistic timelines for the implementation of different areas of AI research? How may the academic-industrial interface vary between the nations that would bring forth separate difficulties in implementing AI research? How will these research agendas be implemented? How should the information from this workshop best be disseminated?

These questions guided the design of the four main thematic topics with which we used to both organize the workshop attendees into groups and to curate the outcomes. The thematic topics also serve to structure the conclusive report here.

1. **Two-Year Horizon Programs:** What can be funded now to make a scientific or societal impact in AI in the next two years?
    a. What AI challenges are important to solve immediately?
    b. What potential breakthrough in AI could we make if we just had more funding in the next two years?

   c. What AI challenges can we solve or make significant progress on in the next two years?
2. **Long Term Programs:** What are important directions for AI research for the long-term future?
   a. What deep scientific questions should the AI community tackle?
   b. What potential breakthroughs in AI could we make if we had long-term sustained research funding?
   c. What are problems that are best or necessarily addressed by academia, those that go beyond the time horizon in industry?
   d. Given the relative strengths of the US and UK, what should we be working on that adds maximal benefit to both nations by drawing on the strengths of each?
3. **Big AI and Small AI:** Many recent advances in AI have relied on Big Compute (e.g., massive GPU clusters) and Big Data, and as such, have largely been developed within industry.
   a. Is “Big AI'' here to stay? If so, how can academia participate and contribute, and what is the role of government research funding? What then is the role of academic AI research, given that academia can and should think further ahead than industry can today?
   b. Assuming Big AI is here to stay, what is the equivalent of the Large Hadron Collider for AI? With government support, should we build one (or two)? How can we ensure it is an open and shared facility? Or, should new models of engagement between academia and industry, perhaps incentivized by government funding, provide an alternative approach?
   c. Looking ahead, is there the equivalent of ``the end of Moore's Law'' for AI, where adding ever more compute and more data will yield diminishing returns? Will AI inevitably be a Big Science, as astronomy and biology are?
   d. Even if AI is a Big Science, is academia's niche to focus on Small AI, especially as we want to reap the benefits of AI on small data and/or small compute (e.g., resource-impoverished devices at the edge)?
4. **Increasing and Diversifying Talent:** How can we better “democratize AI” (US terminology) or “level up in AI” (UK terminology)?
   a. What programs can the US and UK fund to cultivate more talent in AI, especially those in traditionally underrepresented and underfunded areas (geographic, gender, race, socio-economic)?

b. Are there any opportunities for collaboration between the US and UK in increasing the talent pool at all educational levels?
c. How can we best facilitate the porosity of people (students, postdocs, faculty) between nations? What logistical, administrative, or governmental impedances can we improve?

The workshop consisted of two 4-hour events on sequential days, and was attended by over 40 researchers from the US and UK (see Appendix).

The first day had an introductory session with welcoming remarks, two keynote speakers, and a Q&A session. This introductory period was followed by a breakout session where the workshop cohort was divided into four groups to discuss one of the four main thematic topics. Following this breakout session, the attendees re-grouped to discuss their findings.

The second day was similar, but without the introductory session. The cohort was again divided into four breakout groups to discuss one of the thematic topics and then reconvened to discuss their findings. Here, we made sure to have every attendee join a new breakout session to discuss a different topic from the previous day. The workshop concluded with a final writing oriented breakout session where the attendees were asked to formally write out their recommendations for action in long format and then present them to the group in a final plenary session.

The findings and recommendations from these discussions follow.

Key Findings

Discussions amongst workshop participants were naturally broad ranging. Three high-level takeaways are: (1) There was no "Eureka!" moment, e.g., in identifying a new area of research or in proposing a new kind of talent program to fund; (2) While participants might have been instructed to brainstorm for a two-year versus long-term horizon, few "stuck to the script"; many of the challenges need both a short-term infusion of funds to address a challenge's urgency and a long-term commitment of investments because of its need for systemic change; (3) While we expected discussions to focus almost exclusively on scientific matters, participants spoke about AI's societal impact even when speaking about science. To the workshop organizers, the first two takeaways were not surprising, but with respect to the third, it was a surprise that participants homed in on raising societal issues almost immediately, regardless of what session they were in.

Based on the discussions, we have therefore grouped the key findings of our participants into five areas, as follows:

1. Core AI Science
2. Societal Impact
3. Application Research Areas
4. Compute and Data Infrastructure
5. Diversity & Talent

1. Core AI Science

Naturally enough, workshop participants discussed a range of issues in the core science of AI and ML – for most participants, these issues represent the core of their working lives.

Trustworthy AI

As AI systems find more applications – and more safety critical applications – the issue of trustworthiness becomes ever more prominent. With the participants, discussions around trustworthiness explored several different aspects of the problem.

On the one hand, the problem of trustworthiness is viewed as a manifestation of the classic correctness problem studied in computer science since the 1960s. From this perspective, an AI system is judged trustworthy if there is a formal proof showing the behavior to be as the system designer intended. In mainstream computer science, the most successful approach to formally demonstrating correctness is through formal methods, including SAT solvers, model checking, and proof checkers. These techniques are used in industry, e.g., Amazon, IBM, and Microsoft, for approaches such as proving critical components of hardware and software systems are correct to well-defined specifications. Unfortunately, the core algorithms for these techniques simply are not appropriate or do not scale for ML systems in general, and for neural network systems in particular. While development of alternative approaches for checking ML systems is underway, this work is at a very early stage, with significantly more work required – a research priority flagged by a number of our workshop participants.

Another dimension to trustworthiness relates to the *brittleness* of contemporary neural net systems. Put simply, they can fail in unpredictable ways. Famously, the field of adversarial machine learning demonstrated that changes in inputs to a neural network imperceptible to human observers may lead to dramatic changes in the output of the network, leading, for example, to a network apparently making nonsense classifications[6]. This is deeply troubling, particularly when contemplating the deployment of ML systems in safety critical applications. Thus, much more work is needed to understand issues such as how ML models fail and how these failures are prevented.

A third dimension relates trustworthiness to ensuring that AI systems properly learn human preferences and values. When ML programs are trained in the context of some reward system that we have designed (*reinforcement learning*[7]) there is a real danger that the program will learn some unintended and undesirable behavior. Significant research development is required to ensure that ML programs learn the behaviors that we intend and that what they learn embodies essential human values. The latter facet is ethically problematic, as the question of *whose* values should the program learn is judged by some to be an inherently political issue. Tackling this question will thus require contributions from philosophy (ethics and moral philosophy) and law.

---

[6] https://arxiv.org/abs/1412.6572

[7] R. S. Sutton and A. G. Barto. R*einforcement Learning: An Introduction (2nd ed*). MIT Press, 2018.

Explainable AI

As with trustworthiness, over the past decade the issue of explainable AI continues to grow in prominence. Despite the intense attention the issue receives, it remains essentially an open problem. The specific issue is that current ML technologies – most notably neural networks, but including others such as reinforcement learning – are optimized to do specific tasks, but are not able to give any explanation or justification for the choices they make. A neural net, at its core, is a (very large) collection of numeric values attached to network components derived during network training. Generally, we do not have any specific methodology for going from these weights to any kind of explanation about why a system made a decision in a specific situation. Once again, this limitation makes it more difficult to accept ML solutions in certain domains. In an extreme example, consider the use of AI/ML in legal domains. This is an area where many commentators find the use of AI/ML technologies to be deeply problematic for a host of reasons. An inability to explain a decision is inadmissible in this domain; we cannot begin to contemplate the use of AI/ML techniques in legal domains if such systems cannot justify their decisions coherently, in a way that is open to examination by human legal experts.

The explainability issue presents itself in at least two distinct ways.

First, for specific domains (medicine, law, finance, biology, etc.) accepted standards of explainability within differing domains need to be defined. Even within the same domain, different standards are likely to apply for different problems. For example, considering driverless car technology, an engineer working with this technology will expect a different level of explanation to that of a regular user of the technology. Similarly, within the field of medicine, an experienced physician working with an AI system would have different standards for explanation to those of a patient using such a technology.

Second, more research is required in order to understand the issue of explainability in ML. This science is challenging as it points to one of the main issues in contemporary ML - these systems work, in the sense that they (sometimes) give us good answers, but there is no deep theory that explains *how* they work.

Benchmarking AI

A very recent phenomenon in AI is the emergence of extremely large scale ML systems trained on very broad data sets. In summer 2021, a team at Stanford University coined the term "foundation models" to describe such systems[8]. The hope is that such systems learn a broad set of capabilities that can then be specialized by further training for specific applications downstream. Such models, of which GPT-3 is probably the best known[9], show remarkable capabilities in areas such as natural language understanding and generation, and to some extent, in demonstrating commonsense reasoning and related tasks.

Impressive as they are, the limits of such models are easily found; on apparently simple tasks they can produce nonsensical outputs, while simultaneously producing perfectly sensible outputs on seemingly similar problems. This raises issues of trustworthiness (see above), but also presents us with an extremely unusual problem: we have developed systems that have clearly learned some important capabilities – but *what* exactly these capabilities are is unknown. This means that it is very hard to know when and where these systems can be deployed with confidence.

Therefore, there is certainly a need to derive appropriate robust and usable *benchmarks* for AI, so that capabilities of any given system are precisely quantifiable along a range of dimensions. For example, for broadly capable AI systems like GPT-3, we might be interested in knowing their ability in common sense reasoning, understanding of concepts like space and time, understanding of the theory of mind, linguistic capabilities, problem solving capabilities, and so on. A recurring problem for previous attempts to develop benchmarks is that when a set of benchmarks are proposed, researchers will very naturally try to optimize their systems strictly to those benchmarks, thus rendering the benchmark largely meaningless as a measure of capability. Benchmarks need to be broad and resistant to attempts to game them and developed in different areas, perhaps with different benchmarks for different high impact domains. This is likely to require suitable data sets and challenge problems, as in the venerable RoboCup scenario[10].

Efficient AI

Although a cliche in machine learning, it is undeniably true – the human brain is the most intelligent general system we know of with a power requirement of just 20 watts on average; current machine learning systems require so much compute power that they have raised

[8] https://arxiv.org/abs/2108.07258
[9] https://openai.com/api/
[10] https://www.robocup.org

concerns about sustainability. Machine learning can and should be significantly more efficient in compute resources required to train systems and of the data requirements to train them. The current state of affairs also has the side effect – noted previously – that large ML systems are effectively owned by big tech, who have the data and compute resources to develop them. Moreover, data in some domains is inherently sparse, and for such domains, contemporary ML techniques are of limited applicability. As one participant put it, “Going from checkers to Go in AI took 15 orders of magnitude increase in computation budget. Even one more comparable increase would far exceed any feasible budget for the solar system, and the universe lacks enough data. Is this trend sustainable – will it produce the kind of AI that we require?”

For these reasons, there is an international research agenda in efficient AI[11] – but much more work is required to make ML algorithms compute and data efficient.

Research funders must therefore embrace the importance of research in smaller models, whether with less data or smaller compute resources. Small data is critical for democratizing AI, making AI more inclusive, and engaging in new application areas (many applications start with small data before accumulating large data sets over time). This includes learning from data and human advice. Finally, small models that can support causality and better interpretations are an important area with connections to the other challenges below. There is consensus from both the UK and the US participants that this is an important area where most investments are needed.

Integration of symbolic, neuro, and probabilistic architectures.

AI is a very broad field, embracing a range of different techniques. Neural networks are the predominant AI technology at the time of writing – although, as we noted above, this has not always been the case, with symbolic approaches prominent for many years. At the same time, probabilistic approaches, which draw upon techniques for computational statistics and probabilistic reasoning, also demonstrate their value in many domains.

Many of our workshop participants argued that progress in AI would benefit from research on combining these techniques in a close and integrated way. For example, techniques such as SAT solving and constraint satisfaction techniques developed within the symbolic AI tradition are often extremely powerful when applied to many combinatorial problems. How can we build

[11] See, e.g., https://openai.com/blog/ai-and-efficiency/

AI systems that are able to draw upon these different approaches so that the result is greater than the sum of the parts?

New hardware paradigms

Sometimes apparently rather modest developments in hardware have dramatic effects on AI; the emergence of GPUs for ML is the most obvious recent example. If quantum technologies definitely demonstrate "quantum advantage" (i.e., that they genuinely provide theorized benefits over classical computing) then the effect on AI will likely be *very* dramatic. Opinions are divided with respect to if and when we will see quantum technologies truly working at scale, but there are signs of progress[12], and research into the possible practical uses of quantum technologies in AI/ML seems to be justified. Other novel paradigms (e.g., bacterial computing[13]) may also merit further investigation with respect to AI.

2. Societal Impact

Our workshop participants demonstrated considerable concern about the potential harms made possible by new technologies such as AI, and also demonstrated a desire to address these issues directly.

Facilitating Impactful and Beneficial AI Research

A very strong theme arising in the discussions was the desire to do impactful research – going outside the ivory tower of academia to do work that is of tangible benefit to society. The increasing profile of AI, and increasing awareness of the harms that new AI technologies make possible, propelled "beneficial AI" forward as a high profile topic over the past decade. Many conferences and journals in the field now ask authors to reflect on the possible beneficial uses – and abuses – of their work[14]. The views of workshop participants reflected this trend and noted that these issues are acutely felt by new generations of students, who demand to work on research problems that are not just scientifically interesting and challenging, but which have clear societal benefits. However, the clear view of many participants was that there are

---

[12] https://www.newscientist.com/article/2323540-quantum-computers-proved-to-have-quantum-advantage-on-some-tasks/

[13] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3971165/

[14] https://nips.cc/public/EthicsGuidelines

substantial barriers in the way of academic engagement with such research. In particular, a lack of access to key problem sets and, perhaps more importantly, datasets, was felt to be a major impediment. Some participants also commented that the "publish or perish" culture in academia, and the need to consistently produce research results, created an environment in which the least-risk option for researchers was to focus on "safe" research topics, even if those topics were of limited social benefit.

Concretely, the cohort suggested that research funders and governmental organizations could help shift the emphasis to impactful research by providing (for example) benchmark problems in key areas – realistic problems, but accessible to individual researchers and their groups. Additional areas of proposed focus with key social relevance included climate, energy, traffic, agriculture, health care, food production/supply chains, the urban environment (cities, housing), and manufacturing. The upcoming section on "Application Areas of Research" gives further examples. Many of the participants desired to see these focus areas approached with research funding programs that have a *global* perspective – enabling researchers to work on problems with transnational impact.

Education and Human Centered AI

A great deal of discussion amongst participants in our workshops focused around the issues of human centered AI and the connected problem of education in AI.

Human centered AI is a broad term. Minimally, it means designing an AI system that puts the human in the center, akin to human-centered interaction. More broadly, it addresses the role that AI can, will, and should play in our societies. A key concern in human centered AI is that we need to ensure this powerful new technology is ultimately of benefit to society, recognizing that any such technology will inevitably change society itself. The standard example to illustrate this point is social media. On the one hand, social media has proven to be an extraordinary tool for connecting people – it is easy to find examples where social media has empowered individuals – for example, online global support groups for individuals suffering from a rare disability. But these beneficial uses of social media are not what makes headlines internationally. It is the unexpected consequences of the technology that we are mostly concerned about – the prevalence of fake news being the most obvious example. Human-centered AI, therefore aims to understand how AI technologies will affect societies and individuals. It attempts to identify the potential unexpected consequences of AI technologies, and aims to mitigate potential risks of AI/ML systems, and formulates protocols in an attempt to ensure that AI is used to societal advantage[15].

Discussion relating to education and AI was closely connected with the discussion around human-centered AI, and focussed around three issues:

- First, how do we best educate the general public about the nature of AI, its limitations, potential, benefits, and dangers, so that they are best equipped to make informed decisions in their lives relating to this new technology?
- Second, how do we best support the dissemination of technical knowledge about AI, to meet the needs of industry, commerce, and government?
- Finally, how can AI technologies support education more generally (i.e., in the classroom)?

With respect to the first issue it was felt education programs targeted at specific social/societal dimensions (e.g., dangers of AI on social media, fake news, etc.) were appropriate.

With respect to the second issue, continued investment in support for students, e.g., PhD students, via scholarships is desirable. UK participants pointed to the "Centers for Doctoral Training in AI" initiative, where large cohorts of PhD students were funded in specific areas (e.g., social AI, AI and healthcare, etc)[16]. Each such center is attached to an individual institution, and typically has substantial industrial support. Each center admits ~15 students per annum. Such centers are attractive for universities (who receive funding for substantial numbers of PhD students), for non-academic partners (who have access to a large, focussed pool of talent), and to individual students (who have the support of a cohort working in closely related areas, access to industrial partners, and access to substantial resources attached to such projects).

For AI in education generally, a number of possibilities were proposed: AI for teaching soft skills, for training teachers and counselors via simulators, models for student learning, and AI for measuring learning progress were all proposed as important possible areas for future work.

---

[15] An example of the latter would be the "Turing Red Flag" rule, which says that an AI chat bot should ensure that a human conversation participant should always be made aware that they are interacting with an AI system, rather than a human being.

[16] https://www.ukri.org/what-we-offer/how-we-work-in-ai/ukri-artificial-intelligence-centres-for-doctoral-training/

3. Application Research Areas

Not surprisingly, workshop participants identified a host of specific application areas that they felt would benefit from more effort/funding. Here we list those areas that were most prominent in discussions.

AI and Finance

Financial crime is a huge societal problem. Can AI help us end financial crime, end bias in financial decisions, support financial decisions (retirement, savings), identity management, contribute to finance for social good, and support digital forensics?

AI and Public Health

Most attention on AI in health is focussed, reasonably enough, on systems such as diagnosis of abnormalities on X-ray/ultrasound scans. However, there are also huge opportunities for AI in public health. Examples include forecasting the next pandemic – assisting in disease surveillance in general beyond COVID, AI based disease prediction for disease dynamics, public health resource allocation (e.g., vaccines), maternal health (e.g., maternal mortality rate in the US 3-4 times what it is in Western Europe and growing), HIV prevention, TB prevention, optimizing and managing health in low-resource communities for example via nutritional programs. In general, there is significant attention in AI & Medicine, but AI and public health is a slightly different focus with much less attention.

AI and Natural Sciences

Applications of AI in areas such as biology are increasingly common, but there seems to be much greater opportunity to develop these in all sciences. For example, in addition to standard natural sciences, agriculture tech has significant potential for AI involvement.

AI and Climate Change/Sustainability

Naturally enough, the climate was a key concern for many participants. There are already initiatives in this area[17,18] but clearly room for many more. How can other disciplines leverage AI for achieving zero-carbon emission? The specific topics may include challenges in grid

---

[17] https://www.climatechange.ai/summaries
[18] https://www.gpai.ai/projects/climate-change-and-ai.pdf

operations with massive renewable energy sources, sustainable food/water/energy, and sustainable transportation.

AI for Assistive Living

This is already an established area, but both the USA and UK face the classic developed-nation problem of an aging population and spiraling healthcare costs. Technology enabling those suffering from (for example) dementia to have a better quality of life at home has huge societal benefits, and likely economic benefits.

4. Infrastructure

AI and the Academy

Most academic delegates acknowledged the fundamentally altered role of university-based AI research in the 21st century. At the turn of the century, state-of-the-art AI research was still done by PhD students on conventional desktop computers. Outside of areas such as robotics (which has always demanded significant specialized facilities), specialized AI research facilities were rare. The present situation is markedly different; access to specialized ML compute resources (such as GPUs) and, just as importantly, data sets, are essential to stay at the cutting edge of contemporary AI research. Researchers continue to respond to this new dynamic in several ways. One common route for academics is to take a joint appointment with industry and to forge an agreement that provides students and postdocs with access to the relevant industrial resources. However, some participants reported that such arrangements, which appear to be increasingly common, are beginning to pose problems for universities, who find their faculty are being "hollowed out" in AI and related areas. Another response has been to focus research on areas where industry does not have the time or inclination to work, for example by elaborating the underpinning mathematics and algorithmic foundations of AI.

Overall, there was much debate about how research funders might support universities and individual researchers. Suggestions included:

- Pooled compute resources for university-based researchers.
- Ambitious challenge problems to foster fundamental research and bring groups together.
- Sharing/hosting of data sets.

- Support for development and maintenance of high quality data sets in areas of importance – an "active data warehouse".
- More support in proposals for the compute needs of academics.

The importance of data sharing was repeatedly emphasized during discussions. Most data sharing at present is done on an *ad hoc* basis, relying on community volunteers, with informal standards, and often hosted on servers with an uncertain lifespan. General consensus concluded there is a clear value add provided by initiatives which enable the creation and maintenance of shared resources – most obviously, data sets. One proposed concept of data sharing included the creation of national digital resources (with appropriate consideration given to concerns about privacy and liberty) around areas such as health, transportation, energy, finance and others within the US and UK .

Many participants also affirmed the importance of community building. Some argued that clarity on what we want was seen as potentially better than a top down investment in massive compute resources. Some admired the community building effects of CERN[19] – suggesting that the CERN community may be CERN's most important contribution, rather than any specific CERN facility.

5. Diversity & Talent

AI has a well-documented diversity problem; in both the USA and the UK, AI researchers and developers are predominantly white college-educated males[20]. This raises a raft of issues, which workshop participants discussed at length.

Diversity in Building and Using AI Models

One issue is that any AI developed by such a narrow demographic will inevitably skew toward reflecting the concerns of just that demographic, thereby potentially putting other groups at a disadvantage. This potential bias is widely recognized now by all sectors – academia, industry, and government. It manifests itself most obviously in an AI model's outcomes (e.g., classification, decision, prediction) and can often be traced back to the data used to train and test AI models. If historical data is used, then the models will likely reinforce historical bias.

---

[19] https://home.cern

[20] https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-6.pdf

This issue is being tackled from a scientific perspective, leading to efforts in trustworthy AI, explainable AI, fairness in machine learning, etc., discussed elsewhere in this report. Our workshop participants presented a number of additional ideas to ensure diversity of thought in building AI models, including:

- *AI for Social Goo*d: This theme of using AI for social good came up often throughout the entire workshop. Funding programs focused on this theme could target all levels of the talent pyramid. Students are inspired by problems at the interface of multiple fields, such as problems at the interface of the social sciences (e.g., economics) and computer science. It was argued that such a program would diversify the disciplinary perspectives in building AI systems.
- *Emphasis on data*: There was a discussion on more focus on sourcing data because of the criticality of data used in training AI models, the digitization of data means almost all fields generate and curate data, the data used raises issues beyond the technical such as privacy, policy, and ethics, and because starting with data is an easier path to learning the technology. Students resonate with the idea of teaching data analysis in context, paired with teaching of programming, to empower entering students to reason about scientific and social problems with public policy implications. A focus on bringing data science into the classroom has the opportunity to bring in many more interested people (traditional mathematics education has very little data science).
- *Language models*: Large language models such as GPT-3 are almost exclusively built for languages which have large digital corpuses available. This inherently excludes smaller languages from AI's benefits. Thus, it was proposed we might build models specific to populations (e.g., African-Americans). As one participant stated – "We need to build an Alexa or a Google Home that talks like you.".
- *Ethics*: Involve underrepresented groups in building ethical AI
- *Design*: Involve underrepresented groups in designing AI systems from the beginning. Getting workers and end-users in the co-design of AI programs (e.g., involving factory workers in designing an AI system that will benefit them). Design in this sense includes designing the data, designing the system, and designing the user interface.

Diversity in AI Talent

Another concern is that a narrow AI demographic will inevitably fail to embrace the full range of skills and perspectives that society offers. For these reasons, and many more, participants advocated for a range of measures to address the problem of diversity. In this section, we enumerate challenges, organized by what can be done to address the pyramid of talent and at the institutional level.

Education is a crucial enabler – initiatives to target primary schools, high schools and community colleges were advocated. Several participants pointed to simplistic university entrance requirements (notably math) discouraging engagement from certain groups. "We need to change the program, not the world," commented one participant.

Unfortunately, AI is often perceived as not engaging with real-world practical problems. Many participants suggested that, especially for early grades and for certain populations underrepresented in STEM, emphasizing how AI can be used to tackle societal challenges may be a way to attract a broader base of diverse talent. Several examples suggested possible ways of engaging with different education levels:

- Video series, podcasts, and hands-on activities could inspire and educate young students, especially those in primary school.
- TRY AI[21] runs micro-internships with about five weeks of structured training for high school students with diverse backgrounds.
- Joint US/UK hackathons and competitions. Example successes:
    - Robocup[22]. Brought together international communities. A possible example of what can scale. Robocup could have some synergy for co-investment. In general, competitions are a key element for driving up interest – ones that don't require travel were perceived to reach a wider base.
    - Kaggle[23]. An excellent example of how to generate excitement and scale.

[21] https://www.try-ai.org/home
[22] https://www.robocup.org
[23] https://www.kaggle.com

    - Focus for competitions could be on solving real-world problems. For example, "Are water resources distributed fairly?" and "Are juries representative of the population?"
- Pre-doc programs, especially for HBCUs and HSIs, where "bridge" training can better prepare students for a PhD program.
- Graduate student fellowships, especially if targeted toward people who would otherwise find it challenging to get support.
- Exchange programs: PhDs in the US followed by a post-doc in the UK or vice versa.
- Programs to address the diversity gap at the professoriate level. The NSF LEAP-MPS program[24] might be a starting point for a broader program.

As the result of concerted efforts, measurable progress has been made on gender diversity in AI, though workshop participants strongly argued that more could and should be done. Much discussion focused on doubling down on underrepresented minorities. As such, programs that target segments on the population, cutting across age groups were discussed:

- Girls Who Code[25]. Offers a range of programs (clubs, summer schools), without necessarily the need to meet physically.
- African-Americans and Hispanic groups, especially in rural areas.
- The Bangladeshi community. It was noted that this community is underrepresented (although South Asians more broadly are relatively well-represented).
- Foreign students. Simplifying immigration processes would help diversify the talent pool. In our workshop we witnessed a lot of immigrant AI talent. Unfortunately, workshop participants reported that talented AI students struggle with immigration processes, and that may mean that well-trained students may ultimately choose to go elsewhere or return to home countries as opportunities open up elsewhere (as they clearly are doing, for example in China).

---

[24] https://beta.nsf.gov/funding/opportunities/launching-early-career-academic-pathways-mathematical-and-physical-sciences
[25] https://girlswhocode.com

Diversity of Institutions

There is growing interest to provide research support to a broader base of universities (e.g., community colleges and many HBCUs) as a way to grow the pool of talent in AI and to increase diversity. However, many of these institutions do not have the research infrastructure that research universities have developed over the decades. They are not set up to write large center-scale proposals (e.g., to the NSF NAII program). Funding for programs to build up this research infrastructure would be needed for them to compete fairly and effectively for research funding in AI. Supporting partnerships between these aspiring universities and traditional research universities is mutually beneficial.

This might be a good opportunity for funding agencies to revisit their proposal submission requirements, to make it easier for all kinds of institutions to apply and compete. Funding agencies might consider processes where early feedback is given to organizations during the proposal process, to help organizations succeed in winning awards.

Also noted is that while investing in HBCUs is good, funding agencies need to ensure these efforts scale, to increase the impact of individual investments.

Recommendations

In the list of recommendations, we use S (short), M (medium), and L (long) to suggest a prioritization of funding in terms of a two-year versus a longer-term horizon. Roughly speaking, S stands for the combination of the innate urgency of the topic and the capability by NSF and EPSRC to invest within the next two years; M stands for a priority area but would require NSF and EPSRC to work either with other partners and/or beyond their usual purview; and L stands for an area that needs a long-term commitment for investment and/or a systemic change (e.g., to policy, culture, society).

Focussed Research Topics

We believe there is a compelling case for focused research programs around the following topics, as detailed above:

- Trustworthy and Explainable AI (S);
- Benchmarking AI (S);
- Integration of symbolic, neuro, and probabilistic architectures (M); and
- New hardware paradigms for AI (M), including quantum (L).

Focussed Application Areas

Furthermore, the following research areas should strongly be considered for support in their development:

- AI and Finance (M);
- AI and Public Health (S);
- AI and Natural Sciences (L);
- AI and Climate Change/Sustainability (S); and
- AI for Assistive Living (M).

Small Data, Small Compute (M)

Funding bodies should embrace the importance of research investigating approaches with less data or smaller compute resources (e.g., few-shot learning or distilled models), as well as human-AI interaction approaches for model building. This is an important area for future research. This action also addresses domains where the data is not yet abundant, e.g. many engineering domains, and research areas where considerations such as causal reasoning and safety are important.

a. We propose work on integration of ‘symbolic’ knowledge (e.g., in scientific domains with an established body of core knowledge) and data-driven approaches that enable scaling with large scale experiments.
b. The world cannot sustain the amount of resources (financially and environmentally) needed for research in large models only, which is an important consideration for industry as well academia. Highlighting this as a critical theme for NSF and EPSRC is undeniably important.
c. We propose investigating approaches to build, train and use AI models that use on the order of thousands of dollars of compute resources as opposed to millions or even billions.
d. We also need to explore how to integrate small models with larger models.

Compute Resources (S)

We recommend a higher allocation of NSF and EPSRC funding specifically dedicated to compute resources enabling academics to achieve their research ambitions in Big AI.

a. The resource should be provided both for small and medium sized proposal opportunities as well as programs specifically focused on support of infrastructure. A large number of mini moonshots is preferable to a small number of high risk bets.
b. NSF and EPSRC should also support a joint effort for a shared national resource across the countries that leverages some of the efforts underway (e.g., the Turing Institute and Hartree Centre efforts in the UK and the National AI Research Resource in the US). While investments in the billions of dollars similar to current industry investment is untenable, moderately substantial provisions for a national resource for academics to use as part of their hybrid models for obtaining adequate resources for research (i.e., a combination of cloud credits, local machines, local university clusters and, in addition, now a national resource) is necessary.
c. Allocation of compute resource funding must include efforts to build a community of users of said facilities that provide clarity on what academia needs; this is potentially as important as a top down investment in massive compute resources. CERN is a guiding example of productive international community building.

Data (M)

Data is of paramount importance in training performant AI models in various contexts. There is a need for large, curated, well maintained data sets as well as simulation testbeds (e.g. large scale digital twins that represent the full interactive complexity of dynamic phenomena in the real world) in a wide variety of domains, such as autonomous driving, robotics, law and medicine.

Communities, such as autonomous driving, experimental physics, and aerospace can be consulted to see how such initiatives are organized.

a. Part of this wider effort should focus on how to make industry-created models available to academia and society, and motivate such sharing, for further research at non-prohibitive charges (e.g., for access to models such as GPT3). Nationally-funded centers could be the independent and neutral hosts to enable this accessibility.
b. International team efforts will enable the creation of larger, higher quality datasets with better maintenance and cleaning of datasets while producing new methods for generation of synthetic datasets. This will accelerate current research efforts by individual university research teams. An international perspective will enable a broader approach to developing datasets that address a variety of concerns (e.g., that may arise in different countries as in legal domains).
c. Achieving these aims around data curation and maintenance are a prerequisite for enabling a host of research questions to be addressed, e.g. Work on digital twins in scoped domains and development of causal models.

Exchange Programs for PhDs (M)

We argue for a program for PhDs to promote diversity & movement (perhaps like the UK Centre for Doctoral Training program). Researchers felt it was important to help the EPSRC draft the call to explicitly include diversity initiatives (i.e., insist a set amount, say 50%, of PhDs are from underrepresented groups). Every student on the CDT should spend time in the US (exchange visits). A similar program in the US should be developed, potentially funded by the NSF.

Notes:

- Any such program must recognize the different dimensions of diversity: gender, race, geography, age, socio-economic status, intersectional groups and more

- Different programs will likely be required for different under-represented groups: not one size fits all.
- It is particularly important to grow diversity in the top schools since that has a measurable impact on growing the diversity of the professoriate (at least in the US).

Inclusion in AI (S)

We recommend a program on promoting inclusion across the US and UK to build community. A few examples are:

1) A mentoring program that pairs up junior students from underrepresented groups (1st, 2nd year PhD students) with senior PhD students or postdocs to help them with career planning.
2) Expanding the impact of inclusion groups such as Women in ML, Black in Robotics, Women in CS and Engineering, etc., to the broader (non-diverse) community to help educate and reduce bias and microaggressions and build allies.
3) Paid parental leave, child care benefits across professoriate, postdocs, and graduate school.

Early Stages of Pipeline (L)

We urge the introduction of funding to fatten the early stages of the pipeline. These programs can help not just retain talent, but catch talent who fall in between the cracks.

1) K-12: Program funding for training in the middle school and high school pipeline for minorities
    - Relevant in the US (NSF) but not in the UK (not in scope of EPSRC; usually handled by Dept of Education)
2) Post-bacc: Program funding for post-baccalaureates, but for terminal master's or certificate programs, would increase the much needed AI skills for the workforce of the future; jobs would not require a Ph.D.-level of expertise in AI.
3) Pre-doc: Programs addressing college students to prepare them for graduate school.
    - Concrete proposal: "Prestigious" bootcamps for a diverse pool of college students to help give them the technical background for continued success.

Postdoc and Faculty (S)

We recommend a Postdoc/Faculty fellowship program open to those from under-represented groups with the funding (salary, startup funds, travel, housing "soft funding" add-ons to the fellowship) following the award holder. This would make these scholars sought-after by the top graduate schools in the US, and the Russel group in the UK. The holder may then take up posts in both the US and UK over the lifetime of the award. The candidate can choose which institution to go to, letting the holder decide which would provide them with the environment to thrive in the short and long term. This encourages schools to address the issues that might deter under-represented groups from attending.

Closing Remarks

These are truly remarkable times for AI. The current rate of progress and investments is on a scale never before seen. The scientific and technological developments behind new AI present us with unprecedented opportunities – to benefit society, to advance science, industry, government, and medicine, and, of course, to create products and services that will create wealth. However, realizing these manifest opportunities presents us with challenges – not least, that advances in AI increasingly require facilities that place them out of the reach of universities and the public sector.

Our workshop brought together a truly world-class team of AI researchers from both sides of the Atlantic, with expertise from the entire spectrum of contemporary AI. Our participants drew from academia, industry, and government and unambiguously demonstrated the US and UK's status as world-leaders in this most competitive of fields. They also demonstrated a clear appetite for advancing the state of the art AI – an appetite both for realizing the benefits that AI makes possible, while also addressing head-on the well-justified concerns about abuses and undesirable consequences of the technology. The transatlantic research agenda for AI that they sketched out thus included core AI research, issues of societal impact, infrastructure, and the perennial problems of diversity and talent in AI. They also identified a range of promising application domains.

The findings of our participants and their recommendations, documented in this report, represent an insightful snapshot of the current research challenges facing AI. They also provide a well-argued research agenda for transatlantic cooperation in the years ahead. Focussing on these challenges will help to secure the status of both nations as world leaders in AI research and applications, but also provide the best way forward in securing a truly beneficial AI future.

Acknowledgement

We would like to acknowledge the wonderful support of Christian Collier Domenico Vivadelli throughout the organization of the workshop and in the writing of this report.

Appendix: Workshop Participants

The workshop was co-chaired by Jeannette Wing (*Executive Vice President for Research* and *Professor of Computer Science, Columbia University*), and Michael Wooldridge (*Professor of Computer Science, University of Oxford* and *Director of Foundational AI Research, Alan Turing Institute, London*).

The welcoming remarks and keynote speakers were addressed by representatives from each country. Lynne Gladden (*Executive Chair - EPSRC*), and Joydip Kundu (*Deputy Assistant Director - NSF*) provided welcoming remarks. These remarks and follow-on Q&A session were held by Lynne Parker (*Director National AI Initiative - White House*), and Tom Rodden (*Chief Scientific Advisor - EPSRC*).

Workshop attendees represented a broad range of world-class AI expertise from academia, industry, and government from the US and UK:

- Amy Greenwald — *Brown University*
- Aleksandra Korolova — *University of Southern California*
- Alessio Lomuscio — *Imperial College London*
- Andrew Blake — *Samsung*
- Anupam Datta — *TruEra, Carnegie Mellon University*
- Carsten Maple — *University of Warwick*
- Charles Isbell — *Georgia Institute of Technology*
- Daniela Rus — *Massachusetts Institute of Technology*
- David Barber — *University College London*
- Dorsa Sadigh — *Stanford University*
- Emma Brunskill — *Stanford University*

- Guy Van den Broeck *University of California - Los Angeles*
- Kathy McKeown *Columbia University*
- Katie Atkinson *University of Liverpool*
- Manuela Veloso *JP Morgan Chase, Carnegie Mellon University*
- Mark Girolami *Alan Turing Institute, University of Cambridge*
- Marta Kwiatkowska *University of Oxford*
- Michael Jordan *University of California - Berkeley*
- Michael Littman *Brown University*
- Milind Tambe *Google, Harvard University*
- Mirella Lapata *University of Edinburgh*
- Neil Lawrence *University of Cambridge*
- Pascal Van Hentenryck *Georgia Institute of Technology*
- Paul Newman *University of Oxford*
- Peter Flach *University of Bristol, Alan Turing Institute*
- Peter Norvig *Google, Stanford University*
- Philip Thomas *University of Massachusetts - Amherst*
- Pushmeet Kohli *Google DeepMind*
- Reid Simmons *Carnegie Mellon University*
- Somesh Jha *University of Wisconsin - Madison*
- Stuart Russell *University of California - Berkeley*

- Subramanian Ramamoorthy — *University of Edinburgh, Alan Turing Institute*
- Suman Jana — *Columbia University*
- Tom Mitchell — *Carnegie Mellon University*
- Tony Cohn — *University of Leeds*
- Wendy Hall — *University of Southampton*
- Zoubin Gharamani — *Google, University of Cambridge*
- Anne Toft — *EPSRC*
- Erion Plaku — *NSF*
- Henry Kautz — *NSF*
- James Dracott — *EPSRC*
- Kathryn Magnay — *EPSRC*
- Kedar Pandya — *EPSRC*
- Liam Boyle — *EPSRC*
- Nina Cox — *EPSRC*
- Liz Kebby-Jones — *EPSRC*
- Rob Hicks — *EPSRC*
- Roxanne Nikolaus — *NSF*
- Vivienne Blackstone — *EPSRC*
- Wendy Nilsen — *NSF*