A Foundational Research Capabilities Report on

New Foundations of AI

2023

11-1-2023

Copyright & Disclaimers

FRC Study Team

In the Research, Innovation, and Enterprise 2025 (RIE 2025) R&D funding tranche, the National Research Foundation (NRF) of Singapore launched a series of Foundational Research Capability (FRC) studies on cornerstones of modern science and technology that deserve attention and investment. One of the identified focus areas for an FRC study was AI.

An FRC study team spent several months assessing global and local AI research trends to identify New Foundations of AI that are key to the future of AI and its impact on the world at large. They conducted workshops to seek inputs from AI experts, multidisciplinary experts, agency heads, industry leaders and the public. External expert reviewers from the NRF Scientific Advisory Board (SAB), the Committee of Government Scientific Advisors (CGSA) and the global AI research community were sought. The outcome is this FRC report, which will be shared with funding agencies, advisory panels, research institutions and stakeholders in the wider RIE community.

The objectives of the report are: a) to analyze Singapore's AI research performance by comparing against global baselines to identify potential AI peaks of excellence in Singapore; b) to identify the important and potential high-impact long-term scientific and engineering foundations of AI that Singapore should consider investing in; c) to review key AI fundamental technologies that strengthen Singapore's leading position in AI research by building towards the set of enduring foundation capabilities identified; and d) to present suggestions on promising AI research topics, directions and applications for Singapore that will take it into the future, as well as highlighting potential impacts of the suggested future AI research towards a sustainable economy.

The FRC Study Team comprises:

Yew-Soon ONG (Lead)
President's Chair Professor of Computer Science, Nanyang Technological University, Chief AI Scientist, Agency for Science, Technology and Research, Singapore

Ngai-Man CHEUNG (Member)
Associate Professor and Associate Head of Pillar (Education), Singapore University of Technology and Design

Vanessa EVERS (Member)
Professor of Computer Science, Director of Institute of Science and Technology for Humanity, Nanyang Technological University, Singapore

Bryan LOW (Member)
Associate Professor of Computer Science, National University of Singapore, Director of AI Research at AI Singapore

Ah Hwee TAN (Member)
Professor of Computer Science, Associate Dean (Research), School of Computing and Information Systems, Singapore Management University

Cheston TAN (Member)
Senior Scientist, Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

Thomas YEO (Member)
Associate Professor, Centre for Sleep & Cognition and Department of Electrical & Computer Engineering, National University of Singapore

Hanwang ZHANG (Member)
Nanyang Assistant Professor, School of Computer Science and Engineering, Nanyang Technological University, Singapore

Puay Han TAN (Public Officer)
Director of Infocomm Cluster, Agency for Science, Technology and Research, Singapore

Joel SNG (Secretariat)
Head, Foundational Research Capabilities, National Research Foundation, Singapore

Benjamin TAN (Secretariat)
Head, Foundational Research Capabilities, National Research Foundation, Singapore

Contents

EXECUTIVE SUMMARY

BACKGROUND

Artificial Intelligence (AI) no longer resides solely in the domain of experts, gradually becoming a pervasive feature of the human experience. It is expected to play an important role in keeping Singapore competitive amidst myriad challenges such as changing demographics, rising global competition, and possibilities for malicious actions.

One of the FRC studies launched by NRF to dive deep into cornerstones of modern science and technology is this study of New Foundations of AI. This report aims to assess Singapore's AI research performance, identify long-term foundations worth investing in, review AI technologies that can strengthen Singapore's already impressive position in AI research, and to present suggestions on AI research topics that will help take Singapore into the future.

This report is led by Professor Yew-Soon ONG with the core team credited at the start of this report. A diversity of expert contributors from the fields of computer science, engineering, neuroscience, law, psychology and sociology have contributed to the report. It has taken over a year from August 2021 till January 2023 to prepare this report over multiple workshops, peer reviews and rounds of revision.

NEW FOUNDATIONS OF AI

Distilling from historical definitions, AI can be seen as the science and engineering of intelligent machines. The long-term goal is therefore to develop machines that can think, act and learn by themselves, targeting the greater good of countries, economies, and societies. Being cognizant of the challenges and opportunities in present AI research, it is important to set guidance to those who are developing the technologies to realize those intelligent machines and to create positive impact as a result.

To provide such guidance, this report proposes a set of foundation capabilities that will form the basis of long-term endeavours to realize the full potential of AI in Singapore and beyond. These foundation capabilities (which are discussed in greater detail in the New Foundations of AI section) are conceived with the strategy of balancing AI risks and growth. They are briefly described as follows:

Responsible AI
Human responsibility of AI deployment and the science and engineering of intelligent systems along fundamental human principles and values to ensure human flourishing and wellbeing.

Sustainable AI
The science and engineering of AI in a manner where advancement can be continued in the long run, with an emphasis towards reducing the growing reliance of data, compute and other limited resources to achieve performance gains.

Rationalizable AI
The science and engineering of AI with behaviour that can be explained as if a human had performed the behaviour, with an emphasis on describing the underlying reasoning and the causes and effects of the outcomes.

Synergistic AI
The science of specialized AI components working together in embodied machines, and humans and machines complementing one another, thus simultaneously exploiting the best of all worlds to expand the capabilities of both machines and humans.

These foundation capabilities provide the platform to discuss the R&D topics that will strengthen Singapore's position in AI.

KEY TAKEAWAYS AND RECOMMENDATIONS

Singapore is among the world's emerging hubs of AI research with achievements spanning from academic prowess to practical applications. In a field that includes leading global universities such as Harvard and MIT and world-renowned corporate entities such as Google and Alibaba, Singapore universities score impressively in citation count and field-weighted citation impact (FWCI) for AI scientific publications, with peaks of excellence corresponding to major topics of interest worldwide. In addition, Singapore-based young scientists have been honoured as AI world leaders, with 6 of the top 20 global researchers in 2018 and 2020 based in Singapore[1, 2] according to the biennial rankings published by IEEE. That is 30% of all the top AI researchers in the world, which is a remarkable achievement for a country of Singapore's size.

AI is identified as a key cross-cutting enabler across all strategic domains in Singapore's Research, Innovation and Enterprise 2025 plan (RIE2025)[3]: Manufacturing, Trade and Connectivity (MTC), Human Health and Potential (HHP), Urban Solutions and Sustainability (USS) and Smart Nation and Digital Economy (SNDE). Some of the AI-driven applications showcased in the Singapore healthcare ecosystem include the Doctor Covid chatbot at community care facilities[4, 5] and an AI tool to detect pneumonia via chest X-Rays[6] to serve the nation during the COVID-19 pandemic. In the light of increasing diabetes in Singapore, the JurongHealth Food Log[7] app serves those with pre-diabetes by analyzing food photos to provide nutritional advice, and an AI-enabled system diagnoses glaucoma with 97% accuracy[8]. "Artificial brains" to perform bed assignment for more than a thousand beds[9], and the use of robotics and analytics[10] for logistics, patient interaction and precision medicine[11] have been reported in local hospitals. AI applications for the social good include sensing apps EmojiCapcha, Happy Bird and Betterfly that help children with special needs learn about their emotions[12], and autonomous robots for agile, efficient and sustainable cleaning[13]. Localized AI applications that cater to the Singaporean context include a speech recognition engine that transcribes English, Mandarin and Singlish[14], and the world's first Malay and Tamil Speech Evaluation Systems[15].

It is a significant achievement that Singapore has reached this stage in just a short time. While other countries have poured a lot more resources and have large hinterland pools of talent and history of development, Singapore has eked out an impressive position in the last 5 years and can stand shoulder to shoulder with the other giants despite its small population. This is due to the investment in an educated workforce and R&D that we have made over the past several generations. Singapore's BSc in Data Science and Artificial Intelligence has been ranked among the world's top 10 AI and Data Science Undergraduate Courses in 2021[16], which is a noteworthy achievement as it places Singapore alongside illustrious names such as Stanford, Harvard and MIT. Singapore is the only country outside of the US and UK to have a university listed among the 10 courses that were featured. Several top private tech entities have set up joint AI research labs with local universities, research institutions and R&D centres in Singapore, such as A*STAR's partnerships with Singapore Airlines and KPMG, NTU's joint initiatives with NCS, Singtel, HP and Rolls-Royce, NUS' collaborations with Sea and Grab, SMU's partnership with Microsoft for its Centre for AI and Data Governance (CAIDG) and

SUTD's Memorandum of Understanding with DBS for its Design and Artificial Intelligence (DAI) programme. This is important as stellar academic research and partnerships are pivotal for Singapore's continued success in AI.

Here are a few opportunities to build on this progress.

Research: The current research trend in AI including Deep Learning (DL), Deep Reinforcement Learning (RL), Large Language Models (LLMs), Generative Adversarial Networks (GANs) and Foundation Models, is to continue scaling up the learning models. For this, big infrastructure is a necessity, particularly in terms of data and compute. To stay competitive with other academic giants, tech giants and countries, in the short term, Singapore should consider investing big in world-class compute infrastructure to attract and retain talents that seek to pursue large-scale AI experiments, models and challenges that have impactful outcomes. Such big models are however black-box that are not very rationalizable to humans. In the long term, it is thus worthwhile to invest in the fundamental science of AI towards more sustainable models and algorithms that have much lower compute requirements and are also safe, resilient and rationalizable. Current research in DL should also be complemented with classical AI, which has a rich history of successes. A variety of classical and emerging AI algorithms support applications of practical interest, such as data mining, heuristic search and optimization, planning and scheduling, expert systems, multi-agent systems, evolutionary computation, fuzzy systems, game theory, dynamical systems, and human-AI interaction and collaboration. Also important is the need to look out for technological trends worldwide that could disrupt the norm, so that Singapore can keep itself at the forefront of AI research. One example is quantum computing. As quantum computers approach reality at scale, Singapore needs to be ready with algorithms and systems that can work on these machines. Yet another important opportunity worth exploring is open source and decentralized research collectives. Globally, successful examples of such collectives include OpenAI, Stability and Eleuther. Such initiatives need to be encouraged in Singapore so that Singapore researchers can collaborate beyond borders and tap into the best talent the world has to offer. It would also be wise to develop a set of key performance indicators that go beyond publications and assess the real-world impact of AI. Given the emerging importance of Human-AI Synergy (HAS), a multidisciplinary research centre focusing on this field would be ideal. This can be complemented by other efforts such as conducting deep-dives to road-map milestones for a 5-to-10-year timeframe, exploring new avenues for sustained funding, developing a pool and pipeline of multi-disciplinary talent (computational and behavioural) for HAS R&D, and posting HAS-related challenges to the community to identify and train relevant AI talents in Singapore.

Industries: Today's AI has made its greatest impact in internet businesses, through recommendation engines, computer vision or natural language processing. There are many more opportunities in industrial AI that have yet to be well explored. For example, manufacturing system planning and execution faces a unique challenge of every manufacturing plant being different, needing custom AI algorithms specialized to plant-specific data and/or operational conditions instead of a single monolithic AI system that can be rolled out across thousands of plants[17]. AI-driven drug discovery promises to revolutionize healthcare by using machine learning to analyze and find patterns in vast quantities of biochemical data and by using neural networks to design molecules and compounds. Beyond drug discovery, manufacturing of these drugs can also benefit from machine learning that performs quality control, identifies errors, reduces waste, increases speed and enables predictive maintenance of equipment, just to name a few. Other application areas can be seen in urban planning, such as smart grids that match energy demand and supply, distributed smart value chains, accurate forecasting for renewable energy sources, energy-efficient operation of data centres and agri-tech, digital twins, optimization of urban layouts for sustainability and livability, and industrial pollution reduction for air quality improvement. AI also promises to transform scientific discovery through 'self-driving' experimental facilities, high-fidelity data creation through computational simulation, and the ability to work

synergistically with theoreticians, mathematicians, designers, and engineers. Likewise, there are many opportunities in agriculture and the food system, augmented learning, cyberinfrastructure, computer and network systems, and others. More work is also needed on grounding AI in the real world.

Institutes or Initiatives: To make the most of these opportunities, one avenue worth exploring is to set up new research programmes in Singapore. Institute(s) or initiative(s) to focus on the identified foundations of AI such as sustainable, responsible and rationalizable AI can be explored. Yet another promising area worth investing in is human-AI synergy, emphasizing how AI can augment humans, learn from humans, help us understand human intelligence better, work with other AI components, and form collective intelligence. Application areas and industries strategic to Singapore where AI can bring transformative value should also be explored. There is a particular need for multidisciplinary AI research that is informed by a wider range of cross-cutting expertise from computer science, engineering, neuroscience, law, psychology and sociology, among others. An example of such a programme that exists globally is the Stanford's Human-Centered AI (HAI) institute[18]. A potential focus area of a new multidisciplinary institute in Singapore could be the interplay of cognition and neuroscience with other disciplines. Research initiatives that develop AI solutions tailored for local and regional needs such as language and localized issues would be welcome. This would also involve collecting and building models on data that's sufficiently representative of Singapore's (and the region's) demography, culture, languages, nuances and context.

Education: It would be worthwhile to have more synergistic AI PhD programs that are multidisciplinary, cross-universities, and/or cross-research institutes. AI should be considered a core foundation subject at undergraduate level with curriculum flexibility to ensure that students do not just have expertise in one area of AI (e.g., machine learning or deep learning) while being inadequately exposed to others. Secondary and primary school students can be given a fundamental background in AI as well, so that they are aware of the opportunities and risks of AI at a young age. Another welcome initiative would be efforts to identify and define AI competencies for educators in Singapore so that they are qualified to impart vital knowledge and skills to students at all levels.

INTRODUCTION

LANDSCAPE

From its humble roots as a science-fiction trope, Artificial Intelligence (AI) has powered its way into the mainstream over the past few decades. Many well-known scientists, inventors, futurists, entrepreneurs and experts predict that AI will reach and surpass humans in general intelligence in the next few decades.

The history of progress in AI Research is a rocky one that has included several slowdowns known as “AI winters”, but today, AI is well into an era of rapid progress, large investments and great expectations. This is the era of data-centric AI, with some expecting that algorithms will soon be able to automate many of those tasks that a normal person can do in less than one second[19]. This is thanks to a combination of advanced computers, availability of large datasets, and learning algorithms whose performance keeps improving with increasingly larger amounts of data as input.

The past decade has seen AI becoming a pervasive force for innovation as a result of massive research and development (R&D) efforts. To this end, AI techniques are used in a wide variety of fields including data analytics, recommendation systems, chatbots and robotics. Their application domains are also diverse, spanning across various industries such as manufacturing, logistics, healthcare, education, sustainability, security, finance, human resources and many more.

Recent Major AI R&D Areas and Approaches

Contemporary AI R&D today spans a wide range of areas including *Machine Learning (ML), Computer Vision (CV), Natural Language Processing (NLP), Computational Intelligence (CI) and Distributed AI (DAI) including swarm intelligence, Federated Learning (FL) and Multi-Agent Systems (MAS)*. Besides that, there are other areas including planning, scheduling, optimization, symbolic methods, knowledge representation and reasoning, cognitive architectures, probabilistic methods, and neuro-symbolic methods. These areas reflect the current pragmatic mix of both AI functions (e.g., computer vision, natural language processing) and technical approaches (e.g., ML, heuristic search and optimization) as practiced by researchers. Following these, some of the recent and major advances in AI R&D include *Deep Learning, Deep Reinforcement Learning, Large Language Models, Generative Adversarial Networks, Transformers and Foundation Models*.

In particular, *Deep Learning (DL)* has become a predominant AI approach over the last decade. It is a machine learning approach that gained popularity in the early 2010s and has revolutionized AI since. At its core, DL is about learning how to represent data hierarchically, using many layers of simple computing units called “neurons”. One key difference DL has from previous approaches in machine learning is that DL performs many parallel but simple computations instead of a few but complex ones. A key property of DL that has made it so revolutionary is that its performance improves with more data, in a way that is qualitatively better than other techniques. For example, in much of CV and NLP research literature, state-of-the-art results on most tasks such as object recognition and speech recognition are achieved by DL algorithms. However, as DL has boomed in popularity, so has the awareness of its limitations. While DL looks likely to remain the predominant approach in the short term, researchers are actively working on other approaches that may either overtake DL in the long term, or serve as alternative approaches with favourable tradeoffs, such as greater explainability with minimal reduction in performance. DL has first made its impact in CV, but it

has moved on to make significant differences in many other areas of AI R&D including the advances described in what follows.

A related major advancement of the past few years is *Deep Reinforcement Learning (Deep RL)*. Reinforcement Learning (RL) is a machine learning technique that stems from the simple idea that an AI model's desired outputs or behaviour should be positively rewarded, and undesired ones penalized. The impact of Deep RL was first seen in its ability to play games. AlphaGo has famously beaten the world's best human players in the ancient strategy game of Go. Other Deep RL systems have mastered chess, Atari Games, multiplayer games like StarCraft, those with imperfect information such as poker, and subsequently AlphaZero which teaches itself to master games from scratch. Recently, Deep RL has been taken beyond games and is making its way in challenging applications such as robotic learning and self-driving cars. On the NLP front, the recent 5 years or so have seen a rapid rise of so-called *Large Language Models (LLMs)* such as BERT[20], GPT-3[21] and Wu Dao[22]. The largest of these models have hundreds of billions to more than a trillion parameters, making them feasible only for the most well-resourced organizations to create, train and maintain. LLMs were initially used for simpler tasks of language generation, e.g., completing sentences, creating stories, or answering questions. Thereafter, they have been applied to related but complex tasks such as summarization, report generation and translation. More recently, multimodal versions of these models have been created that simultaneously process image, video and text data.

The excitement of researchers and the public in LLMs is largely because these models are able to seemingly generate "human-like" paragraphs of text, although their limitations also became quickly apparent. These limitations include the tendency to wander off into producing illogical statements after a few sentences or paragraphs, and a poor ability to maintain a consistent narrative or train of thought over paragraphs. Perhaps most crucially, such models have been shown to be inconsistent in terms of reasoning, producing answers or statements that are out of context or even irrational.

Another advance is the rise of *Generative Adversarial Networks (GANs)*[23]. GANs are a type of DL models capable of generating new data (primarily images and videos) that appear as realistic as the original data they were trained on. While the earliest GANs generated images that were grainy or strange artifacts, today's GANs can generate images and videos that even forensic experts struggle to tell apart from real ones. Hence, GANs have given rise to the now-common term "DeepFakes" – which means fake images or videos produced using GANs. At the same time, GANs have also enjoyed positive and productive applications such as producing ads, movies or music videos more efficiently, as well as qualitatively changing the creative process and enabling new modes of creativity. GANs have applications in AI R&D as they can generate more data and expand datasets significantly beyond their original sizes. However, it is important to note that the dominant position of GANs has recently been challenged by emerging models such as Stable Diffusion, Midjourney, Imagen and DALL-E-2. For example, research related to image synthesis has shown that diffusion models achieve better image quality than GANs[24].

Singapore's AI R&D Scene

Singapore as a country fares very well in terms of the quality of AI research publications when compared to others globally. Before we dive into the stats, let us first look at a few key concepts and their definitions. Field Weighted Citation Impact (FWCI)[25] is the ratio of the total citations actually received by the denominator's output, and the total citations that would be expected based on the average of the subject field. An FWCI of exactly 1 means that the output performs just as expected for the global average. More than 1 means that the output is more cited than expected, and less than 1 means the output is less cited than expected. Citation count[26] is the

number of times a research work such as a journal article is cited by other works. Scholarly output[27] defines the total count of research outputs, to represent productivity.

According to Table 1 below, data from Scival (2018-2021), based on Elsevier AI Report methodology, shows that *Singapore ranks first in the world* in Field-Weight Citation Impact (FWCI) for AI publications. Singapore's FWCI is 3.51, followed by Hong Kong (3.39), Switzerland (3.12), Australia (3.08) and the United States (2.89).

Table 1: Scholarly Output, Citation Count and FWCI of top countries for AI research (2018-21)

Country/Region	Scholarly Output	Citation Count	Field-Weighted Citation Impact (FWCI)
Singapore	3,767	78,851	3.51
Hong Kong	3,375	69,287	3.39
Switzerland	2,915	53,777	3.12
Australia	7,894	140,350	3.08
United States	50,816	818,841	2.89
United Kingdom	14,857	234,893	2.75
Canada	9,112	130,286	2.65
Netherlands	3,491	51,508	2.65
Germany	10,990	142,799	2.46
Italy	7,447	85,723	2.28
Iran	6,359	84,388	2.28
South Korea	9,836	119,206	2.25
France	7,091	83,418	2.14
Spain	6,649	77,113	2.01
China	85,996	892,675	1.84

A similar picture unfolds when we look at FWCI in individual AI R&D areas, where Singapore is consistently among the top 5 countries in many areas. From Tables 2-5 below, Singapore fares well in the major areas of AI R&D, with FWCI values that place it among the Top 5 countries for Machine Learning and Computer Vision (#5), NLP and Recommender Systems (#1), Multi-Agent Systems and Robotics (#3), and Speech (#2).

Table 2: Top 10 countries sorted by FWCI in the areas of Algorithms, Computer Vision and Models (2018-21)

Country/Region	Citation Count	Field-Weighted Citation Impact (FWCI)
Switzerland	36,429	3.39
Hong Kong	60,155	2.91
United States	454,091	2.85
United Arab Emirates	11,987	2.79
Singapore	43,611	2.74
Australia	74,980	2.72
United Kingdom	111,955	2.49
Canada	53,094	2.09
Finland	9,442	2.08
Germany	73,116	2.07

Excluding countries with fewer than 1,000 publications

Table 3: Top 10 countries sorted by FWCI in the areas of Semantics, Models, and Recommender Systems (2018-21)

Country/Region	Citation Count	Field-Weighted Citation Impact (FWCI)
Singapore	16,927	3.85
United States	132,090	2.23
Netherlands	8,955	2.01
United Kingdom	28,755	1.95
Australia	16,435	1.89
Canada	14,303	1.79
Italy	12,717	1.78
Germany	20,875	1.68
Spain	11,615	1.56
Saudi Arabia	4,675	1.43

Excluding countries with fewer than 1,000 publications

Table 4: Top 10 countries sorted by FWCI in the areas of Multi-Agent Systems, Motion Planning, and Robots (2018-21)

Country/Region	Citation Count	Field-Weighted Citation Impact (FWCI)
Hong Kong	5,344	1.68
Australia	9,348	1.63
Singapore	4,212	1.56
United Kingdom	10,085	1.53
Canada	6,004	1.52
Netherlands	2,170	1.29
United States	27,916	1.28
Italy	4,342	1.21
South Korea	4,495	1.21
France	4,129	1.17

Excluding countries with fewer than 400 publications

Table 5: Top 10 countries sorted by FWCI in the areas of Speech and Speech Recognition (2018-21)

Country/Region	Citation Count	Field-Weighted Citation Impact (FWCI)
United States	66,660	3.10
Singapore	3,996	2.51
Hong Kong	3,437	2.43
United Kingdom	14,893	2.35
Canada	6,562	2.30
Switzerland	3,226	2.24
Netherlands	2,377	2.14
France	6,761	1.89
Germany	9,149	1.76
Taiwan	3,335	1.71

Excluding countries with fewer than 400 publications

Figure 1 below with citation count on the x-axis and FWCI on the y-axis compares Singapore's four research-intensive universities (NTU, NUS, SMU and SUTD) with other global universities as well as leading global organizations. All four institutions score well in FWCI compared to illustrious global universities such as Harvard and MIT as well as corporate tech giants such as Microsoft and Alibaba.

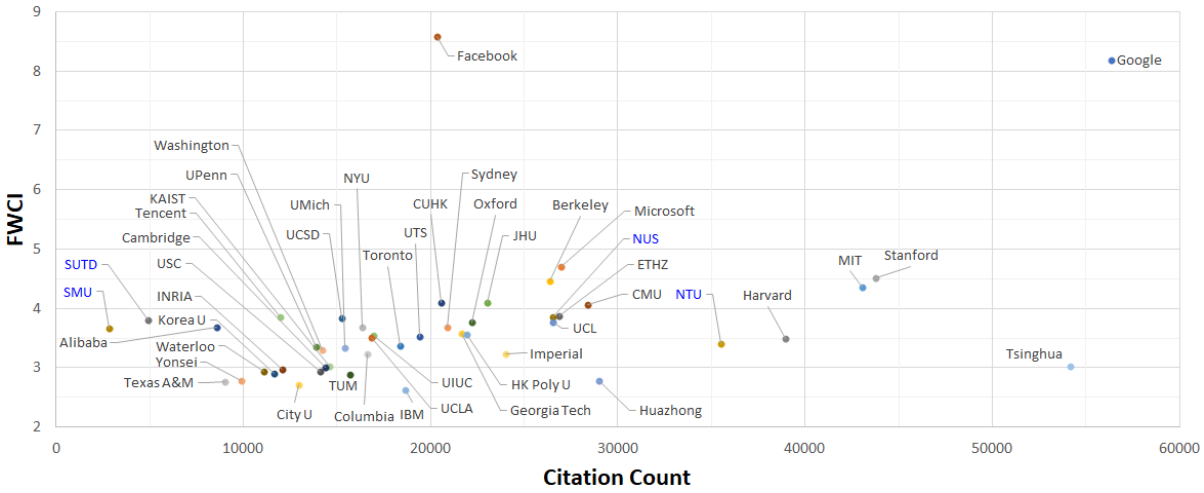

Figure 1: Citation Count and FWCI of top universities and organizations in AI research

There are several current key AI R&D initiatives in Singapore covering both public sector institutions and private Institutes of Higher Learning (IHLs) and Research Institutes (RIs). In 2019, Singapore's *National AI Strategy*[28] was launched to map out how Singapore will develop and use AI to transform the economy and improve peoples' lives. The vision is for Singapore "to be a leader in developing and deploying scalable, impactful AI Solutions, in key sectors of high value and relevance to our citizens and businesses." By 2030. As Singapore embarks on the initial five National AI Projects, it looks at building a vibrant and sustainable AI ecosystem to anchor AI innovation and adoption across the economy. Triple Helix Partnership, one of the five ecosystem enablers identified to achieve this, is where "the Research Community, Industry and Government play important, interconnected roles to enable the development and deployment of AI solutions." It is critical for Singapore to enhance partnerships across these stakeholders to strengthen its collective capabilities and drive overall national AI effort.

AI Singapore (AISG) is a national AI programme to "anchor deep national capabilities in AI thereby creating social and economic impacts, grow the local talent, build an AI ecosystem, and put Singapore on the world map"[29]. AISG is organized into 5 pillars, namely Research, Technology, Innovation, Makerspace and Governance. The AISG Research pillar is a significant funder of basic AI research, via open competitive grant calls; targeted areas include: Collaborative AI, Continuous Learning AI, Privacy-Aware AI, Resource-Efficient AI and Trustworthy and Explainable AI. The AISG Governance pillar promotes AI that is trustworthy and trusted, for the benefit of humanity; its activities include multidisciplinary, applied research grant calls targeting Trustworthy AI and Trust in AI. The AISG Technology pillar addresses major issues that affect the economy and society by organizing various type of applied AI challenges. Challenges have targeted a variety of topics, including healthcare and fake media. AISG also conducts capability-building programmes, such as an apprenticeship scheme[30] and student outreach programme[31].

*The Agency for Science, Technology and Research (A*STAR)* is a national public sector R&D agency that "drives mission-oriented research that advances scientific discovery and technological innovation". A key R&D initiative is the new A*STAR Centre for Frontier AI Research (CFAR)[32]. Led by 2 IEEE Fellows, it focuses on use-inspired basic research, particularly on the three themes of Artificial General Intelligence, Resilient & Safe AI, and Sustainable AI. Another large initiative is the Human-Robot Collaborative AI for Advanced Manufacturing and Engineering ("Collab AI") programme[33], led by A*STAR in close partnership with NUS, NTU and SUTD. It aims to enable a fundamental shift in human-machine interaction, such that machines become capable of working naturally alongside humans as partners. It takes an integrated approach, combining vision, language, haptics, psychology, commonsense and high-level reasoning, orchestrated by a cognitive architecture.

Advances towards AI in manufacturing have also been made by the A*STAR led programme on Cyber-Physical Production Systems – Towards Contextual and Intelligent Response, enabling collaborating computational elements and subsystems to interact with the physical world throughout all layers of a manufacturing production system.

CNRS@CREATE[34] is the first overseas subsidiary of CNRS, the French national centre for scientific research. CNRS@CREATE "acts as a programme operator to build and conduct large transdisciplinary research programmes". The first of such programmes is DesCartes[35], a programme on Intelligent Modelling for Decision-making in Critical Urban Systems. DesCartes "aims to develop disruptive hybrid AI to serve the smart city and to enable optimized decision-making in complex situations, encountered for critical urban systems", where hybrid AI combines machine learning with high-level reasoning.

At the *Nanyang Technological University (NTU)*, the Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU) was jointly established by NTU, Singtel and the National Research Foundation[36]. It aims to "develop applications for use in the areas of public safety, smart urban infrastructure, transportation, healthcare and manufacturing". The R&D efforts focus on AI, data analytics, robotics and smart computing, organized around the three themes of: 1) Anticipatory Analytics and Services, 2) Edge Intelligence and 3) Condition-based Maintenance. Another large AI R&D initiative is the Alibaba-NTU Singapore Joint Research Institute[37], which aims to "make AI become more effective, accessible and inclusive so that it can address future societal needs in ageless aging, new lifestyles and human-centered mobility". Research foci include cloud intelligence, data analytics & intelligence, health AI, AioT technologies and human-centred mobility.

At the *National University of Singapore (NUS)*, the Institute of Data Science (IDS) aims to "leverage and strengthen data science expertise for transdisciplinary and translational research into important real-life problems and education of the next generation of data scientists"[38]. Anchored at IDS, the Grab-NUS AI Lab focuses on mobility-related AI research, research areas include passengers' needs, drivers' behaviours, predicting urban traffic flow, modelling points-of-interest and machine learning for massive transportation data. In 2021, Sea Limited made a gift of S$50 million to the School of Computing to enhance areas such as research in AI and data science. Sea also established Sea AI Lab (SAIL), focusing on a combination of practical and fundamental AI research[39]. More recently, the NUS Artificial Intelligence Lab (NUSAIL) was officially launched in February 2022, focusing on embodied, interactive and trustworthy AI[40].

At the *Singapore Management University (SMU)*, the Living Analytics Research Centre (LARC) "aims to innovate new technologies and software platforms that are relevant to Singapore's Smart Nation efforts"[41]. Research topics include social media listening, multimodal data integration, urban analytics, deep content analytics and personalized recommendations. The Collaborative, Robust and Explainable AI-based Decision-making Lab (CARE.AI Lab) is leading efforts to develop explainable and trustworthy AI systems that can train non-experts to increase their expertise level, especially in safety-critical environments[42]. The Centre for AI and Data Governance (CAIDG), situated in SMU's School of Law, "conducts independent research on policy, regulatory, governance, ethics, and other issues relating to AI and data use"[43]. Its research streams include AI and Society, AI and Business, as well as AI in specific industries, such as autonomous vehicles, finance and dispute resolution. More recently in April 2022, SMU and A*STAR have established the SMU-A*STAR Joint Lab in Social and Human-Centered Computing to "conduct research which integrates social sciences and humanities disciplines with advances in computational intelligence and digital technologies"[44].

At the *Singapore University of Technology and Design (SUTD)*, Artificial Intelligence / Data Science is "an integrated, multi-disciplinary programme which takes a holistic approach to the

education, research and industry implementation" of such technologies, looking not only at near-term solutions but also at fundamental AI challenges, so as to achieve game-changing AI capabilities[45]. There are four research thrusts, namely: theory and fundamentals of AI systems; discovery by AI; human AI interaction; and infrastructure for AI of the future[46]. In line with SUTD's strengths in design, it has developed a first-of-its-kind degree programme in Design and Artificial Intelligence (DAI), aimed at "producing a new generation of designers and innovators for an artificial intelligence-driven world". Possible areas for innovation include better city planning, smarter medical aids and more intuitive digital services.

In the private sector, enterprise software company Salesforce expanded its AI research team to Singapore, its first hub outside Palo Alto[47]. Salesforce Research Asia focuses on machine learning, speech recognition, data mining and business analytics. Other private sector entities have set up joint AI research labs with local universities as mentioned above. A number of other private sector entities have R&D centres or joint labs focusing more on AI development and applications than research, or utilize AI as just part of a broad range of other technologies (e.g., Applied Materials, Delta, Fujitsu, Hyundai, P&G, Rolls-Royce, SAP, etc).

All the above initiatives and achievements reflect a vibrant and exciting AI R&D scene in Singapore, giving immense prospects to consolidate and strengthen Singapore's leading position in AI research. However, given the fast-moving nature of AI innovation, Singapore cannot rest on its laurels. New and disruptive breakthroughs can outflank and outdo all existing accomplishments, or worse, render them obsolete. It is of paramount importance that Singapore keeps aware of the challenges and opportunities in AI research.

Global AI R&D Initiatives

Recent and notable AI R&D initiatives are taking place around the world including the US, China, Europe and Australia. With major countries investing significant amounts of funding into AI, this reminds Singapore of the necessity to continue investing in the field so that it does not get left behind, and instead continues to spearhead progress.

In **North America**, the White House released a Blueprint for an AI Bill of Rights in October 2022 that was lauded as an essential step towards protecting democratic values and civil rights[48]. The National Science Foundation (NSF) established 18 new National AI Research Institutes in 2020 and 2021[49]. The institutes span a wide range of basic and applied research areas. The 11 institutes added in 2021 – after the 7 that were announced in 2020 – cover seven research areas: a) Human-AI Interaction and Collaboration; b) AI for Advances in Optimization; c) AI and Advanced Cyberinfrastructure; d) AI in Computer and Network Systems; e) AI in Dynamic Systems; f) AI-Augmented Learning; and g) AI-Driven Innovation in Agriculture and the Food System. In late 2021, Mark Zuckerberg and Priscilla Chan also announced a sizable donation over the next 15 years, founding the university-wide Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard[50]. Further north, the Canadian Institute for Advanced Research (CIFAR) has established research centers in Toronto, Montreal, and Edmonton, led by world-renowned machine learning researchers.

In **Asia**, the Chinese government has rolled out a series of policy documents and public pronouncements that strengthen AI governance[51]. The Beijing Academy of Artificial Intelligence (BAAI)[52] was established in 2018 to focus on long-term research on the fundamentals of AI technology. In 2021, the institute developed a large-scale language model (Wu Dao 2.0) that reportedly surpassed GPT-3[53]. Wu Dao 2.0 has 1.75 trillion parameters (10 times the 175 billion parameters of GPT-3) and has reached or surpassed state-of-the-art performance on a number of datasets. BAAI has also created the AI computing platform Jiuding which focuses on innovation support, and eVolution which is an intelligent model platform for life sciences[52]. Beyond the work at BAAI, AI in China is being used in enterprise

software, manufacturing (for example using digital twins), transportation (such as autonomous vehicles) and healthcare for rapid drug discovery, clinical trial optimization and clinical decision support[54]. Japan has launched its AI Strategy 2022 whose goal is to realize Society 5.0 (a vision of a smart future) and contribute to the UN Social Development Goals (SDGs) based on the principles of Dignity for People, Diversity, and Sustainability[55]. It aims to achieve these through the five strategic objectives of Human Resources, Industrial Competitiveness, Technology Systems, International Cooperation, and Dealing with Imminent Crises.

In **Europe**, the EU has laid out a common European approach to enable trustworthy and secure development of AI in Europe in full respect of the values and rights of EU citizens[56]. The UK has launched a national AI strategy to leverage AI to increase resilience, productivity, growth and innovation across the private and public sectors[57]. The Alan Turing Institute (the national institute for data science) was founded by 5 leading UK universities (Cambridge, Edinburgh, Oxford, UCL and Warwick) in 2015, with 8 more universities joining in 2018[58]. France launched a national AI strategy in 2018[59] with a focus on three goals: 1) Achieving best-in-class level of AI research by training and attracting the best global talent; 2) Disseminating AI to the economy and society through startups, public-private partnerships and data sharing; and 3) Establishing an ethical framework for AI. Germany also launched a national AI strategy in 2018[60] with a holistic approach comprising: Securing Germany's future competitiveness for the development and application of AI technologies; Ensuring responsible use and development of AI focused on the common good; and Embedding AI ethically, legally, culturally and institutionally through broad societal dialogue and active political efforts. Switzerland is driving AI development in healthcare and pharmaceuticals thanks to its traditional strength in life sciences, political and economic stability that enables safe data storage, and some of the world's best technological universities such as École Polytechnique Fédérale de Lausanne (EPFL) and Eidgenössische Technische Hochschule Zürich (ETHZ)[61]. Microsoft has partnered with Basel-based healthcare giant Novartis to create the Novartis AI Innovation Lab[62]. Google has a dedicated Machine Learning research group in Zurich that focuses on the three key areas of Machine Intelligence, Natural Language Processing & Understanding, and Machine Perception[63].

In **Australia**, the government launched the National Artificial Intelligence Centre in 2021[64]. Established within Data61, the data science arm of the Commonwealth Scientific and Industrial Research Organization (CSIRO), the National Artificial Intelligence Centre includes four AI and Digital Capability Centres. It works to coordinate Australia's expertise and capabilities and build a strong, collaborative and focused AI ecosystem. It achieves this by bringing together partners from government, industry and research to boost AI exploration and adoption in the country, including cultivating a job-ready AI workforce and making it easier for SMEs to adopt and develop AI and emerging technology.

DEFINITION AND GUIDING PRINCIPLES

There have been several attempts to define AI in the past. Some have defined it as intelligence demonstrated by machines as opposed to natural intelligence displayed by animals and humans[65]. Others have described AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated[66].

Existing definitions include: 1) "The science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."[67], 2) "The designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment."[66], 3) "AI as a computerized system that exhibits behavior that is commonly thought of as requiring intelligence."[68], and 4) "Machines that perform tasks normally requiring human intelligence, especially when the machines learn from data how to do those tasks."[69]

Other international research institutions, AI reports and companies have defined AI as 5) "AI is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."[70], 6) "AI as a branch of computer science that studies the properties of intelligence by synthesizing intelligence"[71], 7) "… a set of computer science techniques that enable systems to perform tasks normally requiring human intelligence."[72], 8) "… the science of making machines smart."[73], and 9) "… anything that makes machines act more intelligently."[74]

Below is a definition of AI that is distilled from the earlier definitions put forth and adopted in this reported by distilling from the earlier definitions by renowned academics, researchers, scientists, governmental bodies and big technology companies:

> *Artificial Intelligence (AI)* is the science and engineering of intelligent machines. It allows the machine to perceive the environment, interpret the perceived data, synthesize new information, reason on or process that information, and take actions that affect the environment with some degree of autonomy, in a responsible and sustainable manner.

It is deemed worthwhile to be cognizant of, and to manage, both the pushforwards (exciting possibilities such as AI reaching and surpassing human intelligence in the next few decades) and the pushbacks (risks such as AI redesigning itself uncontrollably or AI being misused by malicious human actors). Singapore is at an exciting juncture in its AI story where it can take stock of its current achievements, relook at research foundations, and propose new directions, opportunities and capabilities.

THE NEW FOUNDATIONS OF AI

With the aim to strengthen Singapore's leading position in AI and drive it into the future, this report proposes 4 new foundation capabilities of AI R&D, namely: *Responsible AI, Sustainable AI, Rationalizable AI and Synergistic AI*. These foundation capabilities are conceived with the strategy of balancing AI risks and growth. These capabilities put forth a new vision for fundamental AI research and innovations that will bring about significant positive impact to the country and the world at large. They will help identify promising future research topics, directions and applications for Singapore. For ease of memory, they form the acronym of RESURG which rhymes with Research.

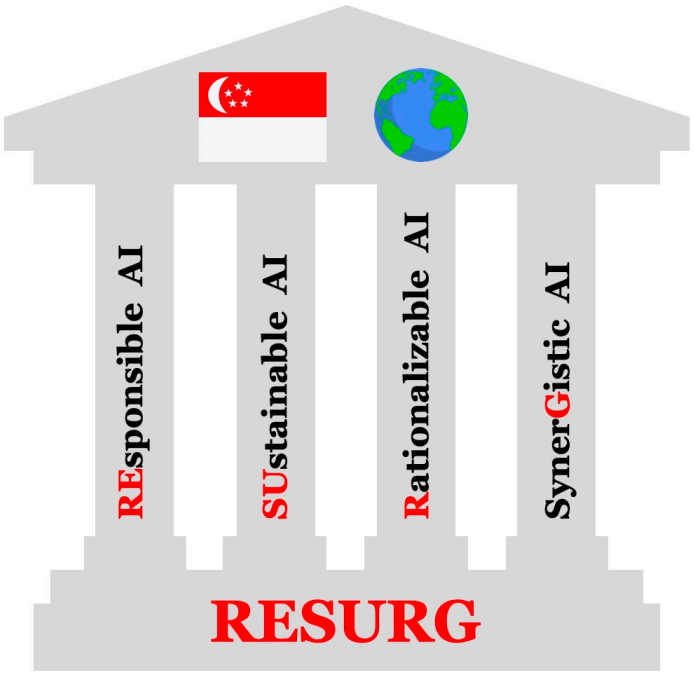

Figure 2: The four identified foundation capabilities

The four new foundation capabilities are *approach-agnostic* in that they make no assumptions that Deep Learning will continue to be the predominant technical approach, and *application-agnostic* in that they will be generally applicable to most eventual applications of AI. Furthermore, they are about long-term foundations and desirable values upon which everything else will be built for the next few years and more.

With the aforementioned foundations in mind, the key AI research topics will unveil not only the challenges of contemporary AI R&D, but also the opportunities that Singapore should consider investing in so as to stay competitive and contribute to the society at large. The subsequent sections shall discuss each of the identified new foundations of AI research in greater detail – looking at related challenges and opportunities while highlighting their importance that led to the definitions of each foundation as well as the gaps to fill through fundamental research.

RESPONSIBLE AI

Responsible AI is the human responsibility of AI deployment and the science and engineering of intelligent systems along fundamental human principles and values, to ensure human flourishing and wellbeing.

There are a few overlapping concepts within Responsible AI and it is not possible to delineate very clearly where one ends and another begins. They could be distilled into the five qualities of trustworthy, safe, resilient, ethical and humane. AI needs to be *trustworthy*, in other words, based on both technical trust (available, reliable and secure) and governance (transparent and accountable). AI needs to be *safe*, in other words, it needs to preserve privacy, data security and user confidentiality. AI needs to be *resilient*, in other words, maintain high predictive accuracy even in the face of adversarial attacks. AI needs to be *ethical*, in other words, avoid biases on the basis of race, gender or any other factor even when faced with biases introduced by data or algorithms. Last but not least, AI needs to be *humane*, in other words, be aligned with human values, enhance human capabilities, empower individuals and society as a whole, and respect human autonomy and self-determination. There are no clear boundaries among the five, with some of the characteristics that make AI ethical also making it safe, and some of the features of trustworthy AI also a key part of resilience.

AI systems are still not sufficiently trustworthy, safe, resilient, ethical and humane in high-stakes applications such as transportation and medical decision support where mistakes have life-or-death outcomes[75]. A malicious actor can simply paste a few pieces of tape on a stop sign to fool a self-driving car's AI into mistaking it for a speed limit sign and continue driving past a stop sign, causing life-threatening risks at intersections. Self-driving vehicles have been blamed for motor accidents[76] in some cases and been recalled due to safety concerns[77] in other cases. In the medical space, IBM's Watson supercomputer was reported to have recommended unsafe and incorrect cancer treatments[78]. The datasets used to train AI may also contain real-world biases that are perpetuated by the AI model to exacerbate racial and gender gaps. There is a wide range of ways in which AI-based recommendation, prediction and decision support systems can exhibit bias or unfairness, caused by the nature of the dataset, the nature of the algorithm, or both[79]. AI systems which exhibit bias or unfairness have been found in many applications, ranging from loan approvals in finance to medical diagnostics[80] in healthcare.

There are also well-founded concerns about autonomous weapons and implications for warfare both in conventional terms as well as cyber-attacks[81]. In recent years, there are increasing negative consequences of AI in social media. AI algorithms have created social media "filter bubbles" by feeding people information they strongly respond to, causing the recirculation and amplification of information that can be extreme, sometimes dubious, outrageous and even shocking, leading to people becoming indoctrinated with misinformation and being converted to radical causes. The human effort needed to generate misinformation and disinformation is substantially reduced as malicious actors have an opportunity to abuse generative text models[82]. On a smaller scale, generative text is being used by students to produce AI-generated homework assignments, thus potentially limiting the students' creativity[83]. In her Harvard University Commencement Speech in May 2022, New Zealand Prime Minister Jacinda Ardern spoke about how algorithmic processes that make choices and decisions for users can radicalize those users, which is why there is a pressing and urgent need for responsible algorithm development and deployment[84].

Globally, the importance given to Responsible AI is evident in how multiple governments such as those of the USA, Canada, EU, UK, Australia and China have rolled out AI governance efforts as we have discussed in Global AI R&D Initiatives above. In addition to governments,

global tech giants such as Microsoft[85] and Google[86] have also laid out policies and principles to ensure responsible AI development and implementation. For example, Google has declared that they will not design or deploy AI in (a) technologies that cause or are likely to cause overall harm, (b) weapons or other technologies whose principal purpose is to cause or facilitate injury, (c) technologies that gather or use information for surveillance violating internationally accepted norms, and (d) technologies whose purpose contravenes principles of international law and human rights.

Singapore has been a frontrunner in Responsible AI. In 2018, the Singapore government initiated an AI and ethics council[87] to address three major categories of challenges for the AI-enabled digital economy: technology challenges (countering data misuse and rogue AI), social challenges (building trust between agencies, companies, employees, and customers), and economic and political challenges (securing Singapore's future in a digital economy). Singapore also established an Advisory Council on the Ethical Use of AI and Data[88] in 2018, comprising representatives from Google, Microsoft and Alibaba, advocates of social and consumer interests, and leaders of local companies who are keen to implement AI responsibly.

In 2019, Singapore introduced its Model Artificial Intelligence Governance Framework at the World Economic Forum (WEF) in Davos. The Personal Data Protection Commission (PDPC) (2019) released its edition of the Model AI Governance Framework to guide organizations to address key ethical and governance issues when deploying AI solutions[89]. The examples of its successful implementation include DBS, Ngee Ann Polytechnic and Visa Asia Pacific among others[90, 91]. In 2020, the Singapore Computer Society also released its "AI Ethics and Governance Body of Knowledge" (AI E&G BoK, 2020)[92] to aid responsible adoption of AI by providing a reference guide on the ethics related to the development and deployment of AI technology. The AI E&G BoK adopted a topical approach to responsible and ethical AI governance, covering key areas such as internal governance structures and measures in AI development or deployment; human involvement in AI-augmented decision-making; operations management in AI development and deployment; stakeholder communications and interactions; and how to get started on AI Ethics implementation, among others. Singapore IHLs have pledged to incorporate the AI E&G BoK into a Mini-Masters in AI and AI Ethics[93]. Singapore is also seeking to develop international standards of AI Governance frameworks and actively collaborates with the U.S. Department of Commerce on this. Bioethics Advisory Committee Singapore is currently working on developing the framework on Responsible AI and big data in healthcare. In addition, the Monetary Authority of Singapore (MAS) has established the guiding framework for Fair, Ethical, Accountable and Transparent application of AI in decision-making processes in the provision of financial products and services[94]. Singapore also has good expertise in formal verification such as the AI governance testing framework and toolkit AI Verify that was launched at the WEF in 2022[95].

A review of research and publications about responsible AI in Singapore uncovered 103 such articles with affiliation in Singapore from 2018 until the present day. The majority of these are articles where AI was used as a tool. Only 9 were articles where Responsible AI was the main topic. Globally, during the same period, there have been 150 articles about Responsible AI, out of which, 70 had responsible AI as the main topic. Furthermore, globally, 47% of the total articles on AI were related to Responsible AI, whereas this was 9% in Singapore. This indicates that research on Responsible AI in Singapore requires more attention, but also shows that other areas of AI are being researched strongly. However, the analysis of published works does not provide a full overview of current efforts. It would be more insightful to understand whether research on Responsible AI is being rewarded in Singapore. There are several recent initiatives in Singapore demonstrating that such research is being rewarded through funding and government support, including the development of Hybrid AI for smart decision-making in critical urban systems[35] and an effort to build a framework for developing certifiable AI systems systematically[96].

This bodes well for the status of Singapore as a smart nation that provides economic and social opportunities for its people in a responsible way. It also positions Singapore as a potential world leader in ensuring trustworthy, safe, resilient, ethical and humane innovation and implementation of AI all over the world. Singapore has a significant opportunity to continue making strides and strengthening its leadership position in Responsible AI.

SUSTAINABLE AI

> *Sustainable AI* is the science and engineering of AI that ensures that its advancement can be continued in the long run, with an emphasis towards reducing the growing reliance on data, compute and other limited resources to achieve performance gains.

Sustainable AI has two distinct components: Sustainability of AI (reducing the environmental impact of AI itself), and AI for Sustainability (the deployment of AI to achieve wider sustainability goals). We will discuss both in this section.

Sustainability of AI is receiving increasing attention as the impressive advances and rollout of AI come with environmental challenges. The computational power and data required to train AI models is already huge, especially for deep learning, and today's trends suggest impending increases in energy demands over the next few years. For instance, state-of-the-art AI models' sizes have ballooned more than 300,000 times in a period of roughly 5 years from around 2012 till 2018[97]. This entails an enormous amount of greenhouse gas emissions that can exacerbate climate change. Google's AlphaGo Zero, for example, the AI that plays the game of Go against itself to learn, generated 96 tonnes of carbon dioxide over 40 days of research training[98]. This is equivalent to 1,000 hours of air travel as reported. These considerations provide an immediate need and motivation to build sustainability into every aspect of innovation and progress in AI so that AI doesn't inadvertently become an unsustainable behemoth with more negative than positive impacts on the planet.

There has been limited progress on techniques that work and score well with low data, models that are computationally compact, and machine generalization techniques that can transfer or adapt knowledge amongst models. These are less carbon intensive models and techniques, but they do not garner as much attention as much of today's AI that achieves high performance at the expense of high computational demand that may only be affordable to major tech giants with ease of access to large-scale supercomputers and AI accelerators.

A recent research area that stands to improve the Sustainability of AI is Tiny Machine Learning (TinyML). It aims to make DL more efficient by requiring less compute, less data and smaller teams to facilitate the emerging field of edge intelligence[99]. Edge intelligence combines AI with edge computing where data storage and computational tasks take place at locations (known as the network edge) close to devices that need them (known as edge devices). Deploying ML algorithms at the network edge allows rapid access to enormous real-time data generated by edge devices, achieving faster AI model training and inferencing[100]. This stands to greatly reduce computation, memory consumption and training costs while only minimally reducing accuracy.

AI for Sustainability is about how AI can achieve positive impacts on the planet. The UN's Sustainable Development Goals (SDGs) provides a reliable set of measures to help us discuss these positive impacts. There are a total of 17 SDG goals (each with a list of individual targets within them) that fall under the pillars of Environmental, Economic and Social sustainability[101]. A recent study[102] pointed out that AI has the potential to deliver positive impacts on 93% of the Environmental SDGs, for example, by enabling smart low-carbon cities and optimizing energy consumption. This was 82% for the Societal SDGs and 70% for the Economic SDGs. Overall, the majority of SDGs in all categories are expected to benefit from AI.

The Oxford Initiative on AI×SDGs, launched by Said Business School in 2021, explores how current and future AI can be used to help support and advance the achievement of the SDGs. The AI for Good Foundation, established in 2015, aims to advance the achievement of SDGs by coordinating AI research communities, policy makers and the general public. The AI4SDGs

Think Tank offers an online open service for everyone, a global repository and an analytic engine of AI projects and proposals that impacts UN Sustainable Development Goals, both positively and negatively.

An example of how AI can power the SDGs can be seen in an initiative in which machine learning capabilities of the SAP Leonardo platform were used for water and wastewater management systems in low-income communities[103]. By predicting water pressure and simulating increases and decreases of pump power, optimal levels of pump power were achieved to ensure that sufficient water was provided to households while not wasting power or breaking water pipes. This powered the "Clean Water and Sanitation" SDG. Another example is from the non-governmental organization (NGO) Elephants, Rhinos & People (ERP) which used AI to analyze photos and videos from unmanned aerial vehicles (UAVs) and camera traps to identify threats to wildlife, for example by detecting humans and identifying vehicle license plates[103]. This facilitated the "Life on Land" SDG.

According to a 2021 WEF report on sustainable development[104], companies are using AI to reduce their carbon footprint, optimize the use of natural resources and optimize the usage of AI to reduce AI incurred carbon footprint. Google, for example, has used machine learning developed by its subsidiary DeepMind to reduce the energy use of its data centres by 35%. The research community is also innovating in this space. For example, Sustainable AI could help predict and forewarn seasonal air quality[105] and applications of machine learning could help mitigate climate change[106]. Research has also been done on the ethical, social and legal dimensions of sustainable AI[107] as well as sustainable AI from the point of view of customer protection focusing mainly on AI related policies[108].

Sustainable AI is a field that calls for significantly higher and continued research investment worldwide, including in Singapore. It is critical to the future of AI that it has a net positive impact on sustainability. Singapore should therefore seize the opportunity to drive progress in this under-explored space and innovate new ways to harness AI for sustenance of our local environment and the planet as a whole. Making early progress in Sustainable AI is a unique opportunity for Singapore to become a world leader in this still-nascent field.

RATIONALIZABLE AI

Rationalizable AI is the science and engineering of AI with behaviour that can be explained as if a human had performed the behaviour, with an emphasis on describing the underlying reasoning (explainability) and the causes and effects of the outcomes (causality).

AI systems today have shown themselves capable of superhuman levels of performance in many tasks. This is especially true for AI systems based on deep neural networks (DNN). These are loosely based on the complex biological neural network that constitutes a human brain, which (unsurprisingly) has drawn significant interest as a dominant source of intelligence in the natural world[109]. Just as one example of a remarkable achievement of a DNN, DeepMind's AlphaFold solved the decades-old protein folding problem which was considered one of the greatest challenges in biology[110].

However, DNNs are often criticized for being highly opaque. They are inherently black-box models that do not facilitate the understanding of their decisions[111]. This is because their layered non-linear structure makes them difficult to interpret and draw explanations as to why certain inputs lead to the observed outputs, predictions or decisions. This causes a natural distrust among many stakeholders who stand to interact with AI and experience its outcomes, therefore hindering the applicability of AI in daily life working side-by-side with humans. This is why AI needs to be rationalizable, or in other words, exhibiting behaviour that humans can understand to a large degree. For example, an AI medical system that can explain itself is important to get the trust of physicians and patients for mass market adoption. In the guidelines set out by the European Union (EU), explainability will be a key consideration to approve new medical AI applications.

The study of rationalizability in AI can be traced back to the history of classical AI in the form of rational agency[112]. In the study of AI agency, a rational agent is one that always picks the best decision according to predefined performance measures. More formally, a rational AI system is one capable of making decisions and performing based on the principles of rationality. Rationality is the quality of being based on or in accordance with reason or logic[113, 114]. Just as rationality a critical attribute of human cognition, it is an essential quality of AI systems too[89, 115].

Rationalizability in AI can be divided into the major themes of Explainability, Interpretability and Uncertainty-awareness. These refer to our ability to show the underlying reasoning process of AI systems that lead to their decisions and actions, either through post-hoc explanations of black-box models or using interpretable transparent machine learning models. It also means the extent to which the internal mechanics of a machine learning system can be explained in human terms. This has given rise to the important field of *eXplainable Artificial Intelligence (XAI).* In the guidelines set out by the European Union (EU), explainability will be a key consideration to approve new medical AI applications. Learning theory can play an important role in explainability as it can help us understand whether and how a learning algorithm can achieve the expected prediction performance, when a learning model can be successfully trained, and how fast the algorithm converges. *Interpretability* also has to do with how accurately a machine learning model can associate the causes and effects of AI systems' predictions and decisions. It entails the extent to which cause and effect can be observed within a system, which is studied under the field of *Causality in AI.* The other key aspect of Rationalizability is *uncertainty-aware learning*. Machine learning and deep learning need the ability to handle uncertainty similarly to how humans deal with uncertainties in various aspects of their lives. AI and ML systems need to incorporate uncertainties into their learning and reasoning processes so that the quality and certainty of their decisions can be assessed.

Most of all, the need for rationalizability cannot be compromised in safety critical applications where it is imperative to fully understand and verify what an AI system has learned before it can be deployed. Examples of such applications include medical diagnosis and autonomous driving where people's lives are immediately at stake. A real-life example of how a lack of rationalizability can be life-threatening is a rule-based AI system that learned the clearly dubious rule that a history of asthma causes a *lower* risk of death in pneumonia patients[116]. This incorrect learning was a result of the training data. Patients with a history of asthma who presented with pneumonia were usually admitted to the Intensive Care Unit (ICU) where they received more aggressive care that lowered their risk of dying. Because the prognosis for these patients is better than average, AI models trained on this data formed the misconception that asthma lowers risk, when in fact, asthmatics have much higher risk of dying from pneumonia if not hospitalized.

Therefore, an ability to examine and verify an AI system is of paramount importance. The development of rationalizable models, grounded in established theories, can thus go a long way in protecting against potential mishaps caused by the inadvertent learning of spurious patterns from raw data[117, 118]. Another aspect to consider is the tug-of-war between accuracy and explainability of state-of-the-art AI systems and this trade-off needs to be considered carefully depending on the applications. The lack of established observable metrics and the absence of human-centred explanations also need to be solved. A lack of agreement on vocabulary and definitions may cause confusion among the literature. On a more positive note, a likely growth area in the next few years is neuro-symbolic AI which seeks to combine traditional symbolic AI approaches with modern deep learning techniques[119].

Globally, the last decade has seen a three-fold rise in research papers concerning interpretability[120], while explainable AI is also discussed as an emerging field[121]. Singapore needs to make further efforts to be toe-to-toe with the global landscape. The good news is that Singapore has already made some progress in Rationalizable AI[122, 123] and does not have to start from scratch. Instead, it stands to build on this progress with continued focus and investment.

In Singapore, IHLs and research institutions have conducted research related to XAI. One of these is *NUS and NUHS's Explainable AI as a Service for Community Healthcare*, in which researchers are building a platform to enable the delivery of AI as a service[122]. Another is *NTU-UBC Center of Excellence for Active Living for the Elderly (LILY) and Alibaba-NTU Joint Research Institution (JRI)* that work towards the aim of developing human-centric AI that is explainable, trustable and creative. NTU has also produced work towards medical XAI that explains AI analysis of medical images[124]. The *Collaborative, Robust & Explainable AI-based Decision-making (CARE.AI) Lab at SMU*[42], funded by AI.SG Research Programme grants, has been working towards collaborative, robust and explainable AI decision making, specifically on agent training programs for safety-critical environments and situational assistance.

SYNERGISTIC AI

Synergistic AI is the science of specialized AI components working in embodied machines, and humans and machines complementing one another, thus simultaneously exploiting the best of all worlds to expand the capabilities of both machines and humans.

AI has caught up with (and in some cases exceeded) humans in several areas. For example, AI has achieved human-level performance in tasks such as object recognition in images, and super-human performance in other tasks such as predicting protein folding or controlling a nuclear reactor. AI also has the advantage of never becoming physically tired.

On the other hand, humans continue to be better in out-of-domain contexts, i.e., new situations not encountered or experienced. AI suffers significant performance degradation in such contexts, and also has difficulty modelling the mental states of humans. Moreover, unlike humans, AI is still quite compartmentalized. For example, CV focuses on just one sense (sight), NLP focuses on just listening and speaking. On the other hand, the human brain is outstanding at processing multiple stimuli at the same time. Considering that AI and humans have distinct strengths and weaknesses, it is unlikely that one can fully take the place of the other. This means that it is important to prioritize developing synergistic AI that effectively collaborates with humans as well as other AI to deliver better outcomes in every field it is deployed in.

AI research has yet to successfully create biologically plausible AI that emulates human cognition and human brain circuitry. AI is still lacking in human cognitive qualities, abilities of biological systems, and common sense. It remains unknown what aspects of biological brain circuitry are quirks of evolution rather than key components of intelligence. Efforts to add more biologically-plausible components have led to performance similar to (but not better than) state-of-the-art AI models. It is unclear whether that is because these added biological components are not necessary for intelligence, or whether the implementation needs further key elements.

To enable AI to emulate the human brain's ability to process multiple stimuli, more efforts are needed to assimilate diverse AI components into holistic systems, such as models that can process multimodal sensory stimulus such as visual, linguistic and others. While that is in progress, new ways of working will also have to be refined. AI can augment humans while humans give feedback or correct the AI so that their collective intelligence achieves optimal outcomes. New job roles for humans can be created that exhibit an understanding of how humans and AI can work together.

Human-AI Synergy refers to how humans and AI agents work together to achieve shared goals. Different from pure AI research which relies on AI alone to solve problems, human-AI synergy is predicated on combining human and AI in solutions. It draws on the disciplines of machine learning, computer vision, natural language processing, robotics, human-computer interfaces, affective computing and more to work towards a new era in the development of AI, characterised by trust and understanding between humans and AI.

Another noteworthy trend in human-AI synergy is the rise of AI-generated content (AIGC) created from human prompts. These include text-to-image tools such as Stable Diffusion[125] and NUWA-Infinity[126], video generation models such as Phenaki[127], 3D image synthesis using DreamFusion[128], text-to-speech generation using ProDiff[129], and text-to-video creation using Make-A-Video[130], just to name a few. Such tools stand to work synergistically with humans and reduce human workload and stress in various content creation domains. AI in programming is also an emerging area. For example, DeepMind has created the AlphaCode

AI that can write code to solve arbitrary problems, which may help automate basic programming tasks[131]. Several other AI tools such as OpenAI Codex, Tabnine, CodeT5, Polycoder and Cogram are able to write code to assist programmers[132].

Cognition and Neuroscience is about building biologically-plausible AI that emulates biological systems. This includes emulating human cognition (mimicking behavioural manifestations of human cognitive functions such as attention) and emulating human brain circuitry (mimicking brain circuitry that supports cognitive functions, e.g., recurrent connections). Progress in this space could take us closer to the Holy Grail of AI research which is Artificial General Intelligence (AGI) where an AI is capable of performing just about any task a typical adult human could. Other exciting possibilities include the use of AI models as models of brain function and behaviour to understand diverse cognitive functions such as vision[133], reinforcement learning[134] and language[135], the use of AI to make sense of the tremendous amount of brain data facilitated by emerging neuro-technologies[136 - 138], and the emerging field of brain-computer interfaces (BCI)[139, 140].

Another key aspect of Synergistic AI is the synergy between different AI components working in tandem or otherwise known here as *AI-AI Synergy*. AI is increasingly better at perceptual intelligence. Human perceptual intelligence is humans' ability to see, hear, or become aware of something through their senses and form a single unified awareness. Likewise, Synergistic AI aims to integrate multiple senses, an example being the emerging interdisciplinary field of integrating Computer Vision and NLP[141]. But that is not all. With its learning and pattern recognition capabilities, AI now has perception that goes beyond human senses, and is able to recognize, perceive and predict beyond the biological spectrum and traditional objects. Also important to Synergistic AI is the paradigm shift towards embodied AI[142]. This is AI that processes multimodal data streams to control physical objects and affect the physical world. To this end, transformers have not only begun to show the ability to integrate vision, language and decision-making to imitate the activities of humans[143 - 145], but have also been used as foundational models to control the actions of robotic arms and collaborate with humans[146].

R&D TOPICS

The 4 identified foundation capabilities of Responsible, Sustainable, Rationalizable and Synergistic AI provide a fundamental platform to discuss the challenges and future opportunities of present AI R&D. The set of R&D topics below have been identified from a thorough review of AI research topics that are fundamental to building these foundation capabilities. The topics don't strictly fall under distinct foundation capabilities but draw on all of them to varying degrees. Figure 3 below illustrates the R&D topics. The differing proportions of how much each topic focuses on *mitigating AI risk* vs *growing AI potential* are qualitatively represented by the colour weights of red and blue, respectively.

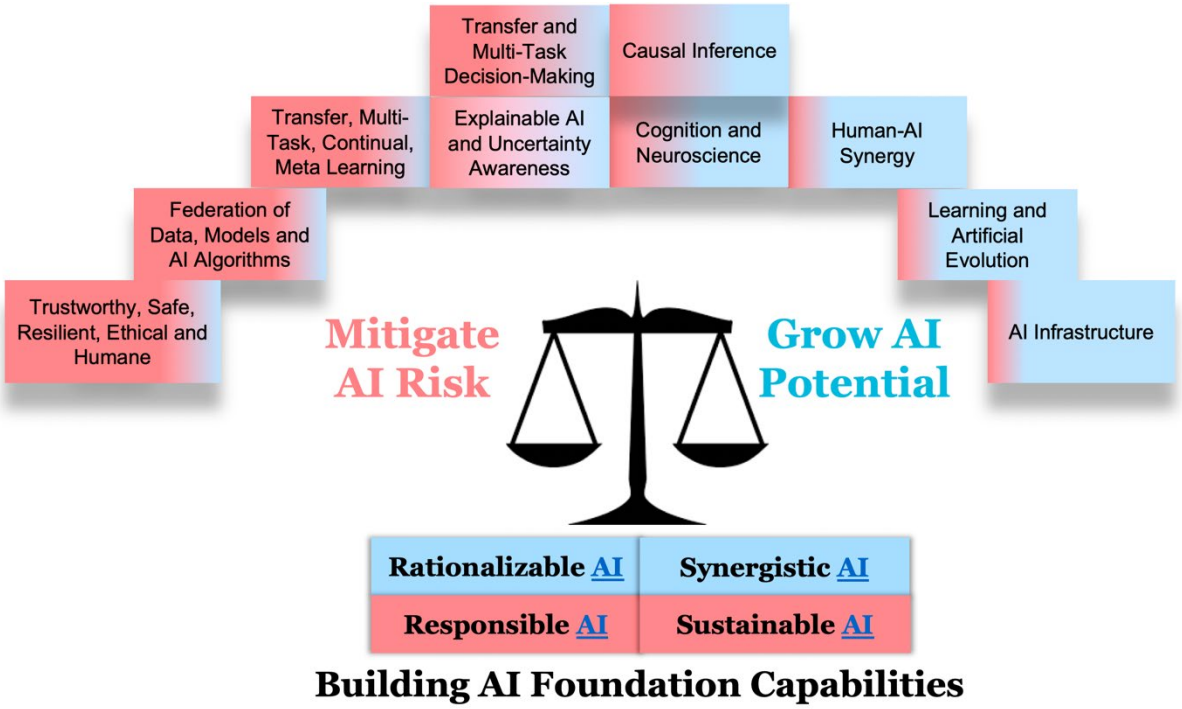

Figure 3: The identified AI R&D Topics

Trustworthy, Safe, Resilient, Ethical and Humane

AI needs to be trustworthy (based on both technical trust and governance), safe (preserve privacy, data security and user confidentiality), resilient (maintain high predictive accuracy even in the face of adversarial attacks), ethical (avoid biases on the basis of race, gender or any other factor even when faced with biases introduced by data or algorithms), and humane (be aligned with human values, enhance human capabilities, empower individuals and society as a whole, and respect human autonomy and self-determination). Singapore should build on its current success in engaging under-represented groups in Tech and AI, focus on augmenting human capabilities with AI and AI capabilities with human, instead of replacing humans with AI, while exploring privacy-preserving learning paradigms such as Federated Learning. Governance frameworks need to be in place to ensure all of these.

Federation of Data, Models and AI Algorithms

Federated learning has emerged as a common framework for distributed machine learning that solves complex real-world problems dependably while preserving data privacy. This is important given the proliferation of IoT devices and regulations such as the Personal Data Protection Act (PDPA) in Singapore. National bodies, research organizations and institutes of higher learning need to set up trusted data, model and algorithm sharing platforms and marketplaces to drive an inclusive and privacy-preserving data economy.

Transfer, Multi-Task, Continual, Meta Learning
Collectively known as machine generalization, these four learning paradigms make use of knowledge or data transfer during training, achieving reduction of computational resources and data requirements, thus delivering sustainability benefits. Collaborative and holistic research for machine generalization is recommended that includes not just mainstream research such as learning algorithms, architectures and models, but also privacy and fairness.

Transfer and Multi-Task Decision-Making
This emerging field involves the transfer and reuse of data, learned models or other experiential priors to achieve rapid search, optimization and problem-solving capability, without requiring vast quantities of data in re-evaluating large spaces of possible solution prototypes. This is important to the sustainability of AI by reducing its computational and data requirements. Continued investment in transfer and multi-task decision-making algorithms is recommended for the automated discovery of compact AI model architectures that uphold environmental sustainability by optimally trading-off multiple performance dimensions such as model accuracy, complexity, and power consumption.

Explainable AI and Uncertainty Awareness
Beyond the accuracy and efficiency of AI's actions, behaviour and solutions, they should also be readily explainable to humans using reasoning and logic that can be easily understood by humans. This is necessary if humans are to entrust AI with influence in their day-to-day lives and especially in high-stakes applications. Explainable AI should incorporate explicit consideration of human and social factors and provides targeted explanation support for specific users and applications.

Causal Inference
To ensure that AI doesn't use spurious correlations to assume causation and make flawed decisions and recommendations, it needs to be upgraded from knowing only correlation to taking causation into account as well, based on the core pillars of causal inference including common sense, imagination and scientific discovery. The development of appropriate guideline metrics is needed to evaluate the results of causal inference. Feature disentanglement, Self-Supervised Learning (SSL), invariance learning, algorithmic recourse, model introspection, scientific rule-regularized learning, and AI for scientific discovery, are promising technical topics related to causal inference. Core statistical methods such as Bayesian reasoning could also be explored to better understand causation [486].

Cognition and Neuroscience
Cognition refers to the study of the human mind and brain, while Neuroscience studies the nervous system and its functions. Together they can help steer AI models and paradigms toward achieving human-level cognitive capabilities and intelligence, thereby taking a step closer to the ultimate goal of Artificial General Intelligence (AGI). At the same time, AI can also help one study and expand the potential of human/animal cognitive capabilities. Continued government funding and support for cognition and neuroscience in AI research are required. Talent development through increased efforts to teach both these topics instead of just one or the other, and the development of open AI-ready testbeds, are important.

Human-AI Synergy
AI has exceeded human performance in some tasks, but humans continue to be better at many other tasks especially those involving new and unseen contexts. As such, synergy between humans and AI can deliver benefits greater than the sum of their parts. This is particularly crucial to addressing Singapore's critical national challenges such as ageing population and labour shortage. Singapore should grow its capabilities in Human-AI Synergy through foundational research into a framework and theories involving multi-disciplinary approaches beyond AI.

Learning and Artificial Evolution
Billions of years of biological evolution can now be efficiently simulated in silico on modern computing hardware, making it possible to replicate in seconds the complex interplay between randomized evolutionary processes and an organism's lifetime learning. One prominent approach is Evolutionary Computation with Baldwinian learning (including neuroevolutionary algorithms) which constitutes a powerful strategy for general optimization intelligence due to its simplicity, flexibility, and ease of implementation Artificial evolution promises to open up myriad new possibilities in AI-AI synergy, such as in generating embodied AI in adaptive robots that interact with one another and with the surrounding environment to perform various tasks. Continued focus on embodied AI by means of synergizing cross-generational evolution with lifetime learning algorithms, leading to the creation of novel robot structures that can operate in various simulated and real environments.

AI Infrastructure
Singapore needs to greatly increase its AI R&D infrastructure to maintain its competitive edge. High Performance Computing (HPC) and data lake are important areas that need immediate investment. Moving forward, disruptive technologies such as Quantum Computing and AI accelerator will be increasingly crucial. A dramatic increase in AI infrastructure is needed while ensuring its sustainability, continued investments in training and skill development of AI talents, investment in emerging technologies such as quantum computing and homomorphic encryption, a unified data policy, and global partnerships.

The next few sections will now take a deeper dive into each of these R&D topics and look at the scientific background in each topic thus far, and a more detailed set of recommendations for building on Singapore's progress in each space.

TRUSTWORTHY, SAFE, RESILIENT, ETHICAL AND HUMANE

Scientific Background

As AI systems play an increasing role in social spaces, they may unexpectedly discriminate on the basis of race, gender, poverty or disability. As an example, a software system called COMPAS used by American courts to judge the risk of an offender recommitting another crime was found to assign higher risk to Black people even when other inputs were similar across races[147]. Such biases have also been observed in facial recognition systems[148] and recommendation systems[149]. This shows that there's work to be done in mitigating bias and achieving social harmony and justice.

Unfairness generated by AI models usually arises from two types of biases: data biases and algorithm biases[72]. The most common data bias is that the data contains sensitive features (also called protected attributes) such as race and gender. The use of such features in AI models will lead to direct or indirect discrimination in some specific tasks. Other biases in data include conformity bias and popularity bias in recommendation systems[150]. Conformity bias refers to how users tend to behave similarly to their friends, and such behaviors do not always reflect their true preferences. Popularity bias means that the popular items tend to be recommended more frequently and less popular items tend to get limited attention.

Different from data bias, algorithmic bias is induced purely by AI algorithms. In ML datasets, commonly used datasets such as IJB-A and Adiance have been shown to be imbalanced, containing mainly light-skinned subjects, thus introducing biases against dark-skinned groups[151]. It was further shown that there's a need to subdivide subjects into light-skinned females, light-skinned males, dark-skinned females and dark-skinned males in order to uncover the hidden biases against dark-skinned females.

Various approaches have been identified to achieve AI fairness[72, 152, 153.] These approaches can be categorized as pre-processing, in-processing and post-processing methods. Pre-processing approaches focus on transforming the data to remove biases and discrimination. Common pre-processing approaches include variable blinding[154, 155], relabelling[156] and data reweighing[124]. In-processing methods include fairness constraints during model training to maximize both performance (e.g., accuracy) and fairness. For example, fairness constraints are considered when optimizing the accuracy for classification[157] and regression tasks[158]. Post-processing methods aim to improve the model fairness by applying transformations (e.g., thresholding[159]) to the model outputs.

Two of the most pressing threats to the safety of DL models are adversarial attacks and data poisoning[160]. For adversarial attacks, the most effective defences are based on a simple concept of adversarial training (AT) where adversarial examples are generated and incorporated as training samples so that the AI models will learn to be robust against them during test time[161 - 164]. Another approach is to minimize the effects of small perturbations on the models' predictions[165 - 167] to train the model to rely on less-superficial features that are more aligned with human vision[168 - 170]. Another class of defenses is provable defenses which seek to provide a performance guarantee for an AI model's performance in the face of adversarial examples[171 - 175].

Data poisoning can happen because modern AI models rely heavily on large amounts of training data, which exposes the models to the threat of attackers who can degrade a model's performance by corrupting a small subset of its training data as a data contributor[176 - 178]. A sophisticated variant of data poisoning called backdoor poisoning allows an adversary to control a model's prediction through a poison signature in the model's input while eluding

detection[179 - 182]. Several defences have effectively countered this threat under certain conditions. One type of defence works by filtering poisoned samples that contain spectral signatures where their hidden states would have different statistics compared to uncorrupted samples[183]. Other approaches include pruning 'suspicious' neurons that lie dormant in the presence of clean validation data[184] or using differential privacy to counter the poison[185]. More recently, provable defence approaches have been studied that can provide guarantees to a model's performance when the data poisoning is known[186, 187].

The defences and attacks on AI systems will likely see a lengthy "arms race", especially in high-stake applications. It is therefore vital to invest resources to develop more advanced defences to stay ahead of the attackers. In addition to safety against adversaries, much work has been done on improving models' reliability in the face of high signal-to-noise environments[188, 189]. Exposing AI models to training samples augmented with a wide diversity of corruptions has shown to improve the performance of models during such challenging scenarios[190 - 192].

Recommendations

Globally, the AI Index Report[193] developed by Human-Centered AI Institute, Stanford University, highlighted low implementation and lack of attention to ethical AI principles. The Montreal AI Ethics Institute's The State of AI Ethics Report[194] highlighted examples of sometimes ambiguous approaches to managing Trustworthy AI by corporates and governments. It also claimed that "public trust in algorithmic decision-making systems is at an all-time low" and recommended that a combination of industry, research, the public sector and the policymakers/regulators is needed to cultivate more trust. A report developed by FICO[195] in 2021 identified that more than three quarters of surveyed participants found it hard to prioritize responsible AI practices, indicating a lack of awareness of the risks imposed by unethical use of AI. Only 35% of respondents took the steps to ensure that they were using AI transparently and with real accountability, and only 39% said their companies were placing greater emphasis on model governance during the AI development process. This means that a majority are still not placing greater emphasis on governance of AI technologies. The Economist Intelligence Unit's "Staying ahead of the curve: The business case for responsible AI" Report[196] sponsored by Google pointed out that lack of clarity on how multiple regulatory frameworks can be applied undermines public acceptance of AI and stalls potential investment.

A pre-emptive approach is important for AI safety. Given how wide and diverse the possible AI applications are, it is critical to scrutinize each stage of an AI system's training and deployment processes for possible entry point where a malicious actor can attack. Federated Learning (FL) could be one way to make AI more trustworthy, as the next R&D Topic will discuss. In terms of how much control an algorithm should be entrusted with, it was discovered that a 60-40% human-AI partnership seems the most plausible as humans would still be in more control[197].

Singapore should continue to represent women and ethnic minorities in its AI research. Singapore is already doing well on this front, outranking US and India in engaging women in Tech and AI. Due to the early involvement of girls in STEM education, Singapore has 28% of females among AI talent pool, having one of the smallest gender gaps in the world. Another recommended focus is augmenting human capabilities with AI[198]. Beneficial human-AI collaboration has already been implemented in Singapore companies such as DBS Bank[199], Changi Airport[200] and Jewel[201]. Integrated deployment of AI would not only redesign social interactions but also enhance human productivity and empower human capital by AI-driven technologies[202]. Further progress in beneficial human-AI collaboration is recommended.

FEDERATION OF DATA, MODELS AND AI ALGORITHMS

Federated learning (FL) is a mode of learning in which multiple devices collaborate to learn a machine learning model under the supervision of a central server[203]. One of its key advantages is that this learning can take place without sharing each device's private data with the other devices. In other words, it is decentralized. In addition to preserving privacy, this also has the advantage of eliminating data communication overhead[204]. The central server aggregates and shares the built knowledge among participants.

The importance of preserving privacy has helped FL become a key framework for distributed ML. FL allows different devices to train on massive amount of diverse, privately-owned and geographically distributed data. Each device builds its own local models whose training processes can be synchronized via sharing differential parameter updates. This can be achieved without exposing their private training data, thus reducing privacy violation risks and staying compliant of laws and regulations such as the Personal Data Protection Act (PDPA) in Singapore and the General Data Protection Regulation (GDPR) in the European Union, for instance. The need for an integrated intelligence that preserves privacy has also been elevated by the proliferation of mobile and Internet of Things (IoT) devices. FL is one promising way to do that. It can enable separate functional components to be synergistically integrated into holistic, intelligent, robust, resilient, dependable and scalable AI systems.

For these reasons, FL is receiving widespread interest from the ML community, resulting in a fast-growing body of studies that evangelize federated learning as the new standard of ML as a service for democratizing AI and establishing trusted data sharing platforms and marketplaces[205, 206, 207]. It has also attracted interest from a variety of real-world application domains like digital healthcare[208], medical imaging[209], wake word detection for smart voice assistants[210], and next word prediction on mobile devices[211], just to name a few.

Scientific Background

FL is significantly different from centralized ML and distributed on-site learning[204]. In centralized ML, data from various devices such as computers, mobile devices and autonomous vehicles are sent to the Cloud, where the ML model is built and then leveraged by a user that sends a request to one of the available services through an API. In distributed on-site learning, each device builds its own model using its local data set, with no more communication with the Cloud after the first interaction that distributes the model to the devices. FL borrows elements of both, such that each device trains a model and sends its parameters to the central server that aggregates it and shares the aggregated ML model with peer devices.

This takes place through 4 steps[203]: (1) *Client Selection/Sampling*: Server either randomly picks participants from a pool of devices or uses an algorithm to select the client; (2) *Parameter Broadcasting*: Server broadcasts the global model parameters to the clients; (3) *Local Model Training*: The clients will simultaneously retrain the models using their local data; and (4) *Model Aggregation*: Clients will send back their local model parameters to the server for and aggregation towards the global model.

The different types of FL frameworks[203] include Vertical FL, Horizontal FL, Federated Transfer Learning, Cross-Silo FL, and Cross-Device FL. FL has been applied in several major domains including healthcare, transportation, finance and Natural Language Processing (NLP). Noteworthy recent developments in FL[203] include One-Shot FL, Incentive Mechanisms, FL as a Service, Asynchronous FL, and Blockchain in FL.

Recommendations

More research is needed to overcome training challenges and security challenges in FL[203]. Training challenges include 1) *Communication overheads* due to frequent communication between the central server and devices, 2) *Systems and data heterogeneity* and non-identically distributed data from the multiple devices, 3) *Barriers to collaborative sharing of data* because data owners may prefer their own pretrained ML models and cannot be sure that the other entities are contributing safe data, and 4) *The need for a more inclusive data economy* that's not just be driven by "Big Tech" corporations with big data but also by government bodies and small-to-medium enterprises and start-ups. Security challenges include *Membership inference attacks* (attacks involving inferring whether certain training data exist from the model information), *Data poisoning attacks* (attacks in which adversaries poison the training data in some of the participating devices so that the global model accuracy is compromised), *Model poisoning attacks* (where local models are poisoned instead of local data), and *Backdoor attacks* (where an adversary can introduce a backdoor functionality into the global model[212]). Addressing all these challenges will be critical to ensuring that FL continues shaping up as a strong standard for privacy preservation in ML.

Trusted data, model and algorithm sharing platforms and marketplaces need to be established that facilitate sharing and/or trading of such resources between owners, while protecting personal data ownership and safety. To achieve this, the following key research topics need to be explored in greater depth: a) Black-box model fusion[213 - 215]; b) Data valuation[216 - 221]; c) Incentives and reward mechanism design in federated/collaborative machine learning[222 - 224]; d) Federated AI planning and sequential decision-making algorithms like federated reinforcement learning[225] and federated/collaborative Bayesian optimization[226 - 228]; and e) Machine unlearning[229 - 234]. These research topics will require non-trivial integration of methodologies, knowledge, and expertise from several AI domains such as ML, data mining, heuristic search and optimization, planning and scheduling, reasoning under uncertainty, multi-agent systems, game theory and economic paradigms, as well as from the emerging domain of AI privacy. Two related research efforts are TrustFUL: Trustworthy Federated Ubiquitous Learning[235] and Toward Trustable Model-centric Sharing for Collaborative Machine Learning[236].

At the national level, joint efforts among national AI programmes, research organizations and institutes of higher learning can garner momentum to achieve greater breakthroughs for privacy-preservation research in AI. For instance, governmental organizations and private companies may consider jointly setting up trusted data, model and algorithm sharing platforms and marketplaces that promote an inclusive data economy by facilitating *government agencies and institutions* with unlocking valuable data from the data vaults, and *small-to-medium enterprises and start-ups* with the federation of their limited data. Major healthcare clusters could coordinate data and algorithms to benefit healthcare. Organizations involved in agri-tech, aqua-tech and food tech could coordinate the federation of data, models, and AI algorithms from the respective farms and companies. At the individual level for instance, governance and implementation efforts towards personal data ownership in data, model and algorithm sharing platforms on mobile apps are recommended.

TRANSFER, MULTI-TASK, CONTINUAL, META LEARNING

Most existing AI systems focus on narrow AI[66, 237], meaning they are designed to learn isolated intelligence to solve specific tasks under specific contexts. While this leads to outstanding performance in these narrow tasks, even small changes in the tasks and contexts would require substantial compute and data overheads to develop the systems. In marked contrast with most AI systems, humans can learn a task with minimal experiences and cognitive efforts. Humans have an outstanding ability to generalize, leverage and transfer prior knowledge to a novel task of different goals and contexts. This strong ability to generalize is a hallmark of human intelligence[66, 237].

Taking the cue, the concept of *Machine Generalization* refers to learning paradigms that make use of *task-level knowledge transfer* during the training of AI models, achieving reduction in computational resources and data requirements. Prominent machine generalization approaches in the AI literature today include Transfer Learning (TL), Multi-Task Learning (MTL), Continual Learning (CL) and Meta Learning (MeL). *Transfer Learning* extracts knowledge from one or more source tasks tackled in the past, and applies this knowledge to related target tasks at hand[238, 239, 240]. *Multi-Task Learning* is the idea learning multiple related tasks simultaneously to achieve computation and memory savings, increase data efficiency, and achieve improved performance in some cases by uncovering common latent representations for these tasks[241, 242]. *Continual Learning* enables a machine to learn a sequence of tasks such that there is no forgetting of previous tasks and there is reduced resource requirement to learn new tasks[243]. *Meta Learning* is about training an AI model by using many similar tasks which are usually assumed to be drawn from a particular task distribution, with the goal of learning a meta-representation that is broadly suitable for many tasks from this distribution, including future unseen tasks[244, 245, 246].

Scientific Background

Figure 4 below illustrates the four machine generalization paradigms in terms of how they perform task-level knowledge transfer.

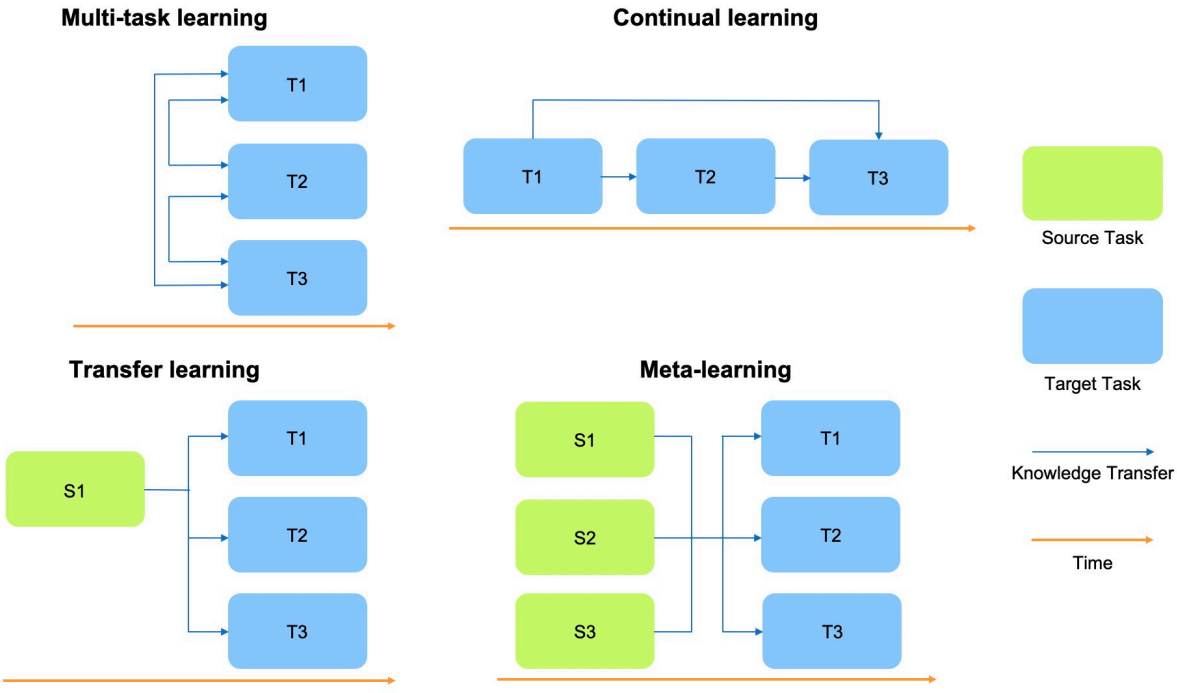

Figure 4: Common machine generalization paradigms

Multi-Task Learning (MTL) provides several benefits compared to single-task scenarios including reduced memory footprint, reduced training and inference time, potential improvement if the tasks are related, and sharing of complementary information[241, 247, 248]. However, a major limitation of MTL is that it only works for related tasks. If the tasks are not related, the notorious effect of *negative knowledge transfer* - sharing the knowledge of a task that is destructive to other tasks – will severely degrade the performance[249, 250].

Continual Learning (CL) has seen many efforts to mitigate the major problem of *catastrophic forgetting (CF)* which is the issue where standard neural networks usually forget most of the information about the previously learned tasks when trained on new tasks[243]. Numerous works have been proposed in the previous years to mitigate CF, and they can be classified into three main categories: replay[251 - 259], regularization-based[260 - 264], and parameter isolation methods[265 - 272]. Recent progress in CL has led to some real-world emerging applications including the algorithms used by Netflix and Amazon[273].

Transfer Learning (TL) has shown an ability to extract and transfer knowledge from source tasks to target tasks in an identical or similar domain[238, 239, 240, 274 - 306]. The training can be viewed as a two-stage process: In the first *pre-training* stage, knowledge is extracted from the source tasks. In the second *fine-tuning* stage, the extracted knowledge is leveraged for learning of target tasks. Pre-training using a large labeled dataset such as ImageNet[307] has empowered learning of many target tasks with limited data. Recent progress in self-supervised learning enables pre-training using unlabeled dataset[308 - 313]. TL has also been applied to generative models very recently to enable training of GAN under very limited data[314 - 319].

Meta Learning (MeL) includes a few different categories. *Metric-based* meta-learning aims to learn a similarity metric between support and query samples by learning an embedding space[320, 321 - 325]. *Optimization-based* meta-learning focuses on learning an optimizer, or finding good initialization points for model parameters for fast adaptation[244, 326 - 330]. *Augmentation-based* methods learn a generator from the existing labelled data to further use it for data augmentation in novel classes[241, 242 - 244]. *Weight-generation* methods directly generate the classification weights for unseen classes[255, 256, 257, 258].

Recommendations

A holistic research direction is important for machine generalization. In addition to mainstream research such as learning algorithms, architectures and models, other aspects including privacy and fairness in the context of machine generalization need to be seriously investigated to ensure successful adoption of machine generalization for sustainable AI systems.

Task heterogeneity has been lacking in some machine generalization problems and setups. As a consequence, AI models trained with minimal elements of task heterogeneity may not be applicable to many practical scenarios where task diversity exists. An example of task diversity could be an intelligent autonomous robot that involves object detection, semantic segmentation, situational awareness and anomaly detection abilities. It is desirable that the robot can learn from a variety of heterogeneous tasks and leverage this knowledge to learn novel tasks with minimal data and computation overhead, achieving resource reduction and learning efficiency improvement.

The human brain is once again to be taken as motivation. For example, a human can learn a new ice hockey trick quickly by generalizing his/her prior knowledge in ice hockey as well as previous skating and field hockey experiences. At present, there is still a significant gap between machine generalization paradigms and this aspect of human intelligence, which needs to be bridged to achieve practical and sustainable AI systems.

While there has been active research in ML privacy[259, 241, 243, 260], different machine generalization paradigms have their particular setups and characteristics, posing new privacy challenges. Research into privacy-preservation in ML has focused mainly on protecting the privacy of individuals whose data was used to train the model[259, 261, 262, 260], but privacy in the context of machine generalization has not been sufficiently studied. For example, in transfer learning and meta-learning, the source models pre-trained on potentially confidential data need to be shared with downstream tasks for knowledge transfer. An adversary may attack the source models directly to extract sensitive information of pre-training data samples, or attack the downstream models which inherit source models' sensitive knowledge. Two privacy attacks are of primary concern in machine generalization. In *membership inference attack*[263, 264], given a data record and access to a model, the attacker aims to determine if the record was in the model's training dataset. In *model inversion attack*[261, 265 - 267], given access to a model, the attacker aims to recover sensitive attributes of the private training dataset.

In existing AI research, data bias and algorithm bias have been the main concerns of unfairness. With machine generalization, a new type of bias associated with task-level knowledge transfer could arise. We refer to such bias as *knowledge transfer bias*, which is associated with knowledge transfer between source and target tasks (for transfer learning and meta-learning) or among individual target tasks (for multi-task learning and continual learning). The bias associated with task-level knowledge transfer has not been adequately studied, which reduces confidence in the adoption of machine generalization in high-stakes scenarios[268].

Overall, machine generalization is an important scientific topic that could have profound impacts on the sustainability of AI research in the long run. In Singapore, strong research capability in this topic has already been seeded in many research institutions. More collaborative and holistic research on machine generalization including learning algorithms, architectures, models, privacy, fairness and other technical aspects would be appropriate.

TRANSFER AND MULTI-TASK DECISION-MAKING

Search and optimization lie at the heart of decision making, which essentially entails deciding what to do to maximize positive outcomes when faced with possibly unknown problems or situations. Importantly, such real-life problems seldom exist in isolation. It is common that similar decision-making tasks routinely recur in industrial applications and everyday life, making the transfer and reuse of experiential priors an essential element of efficient problem-solving. Unsurprisingly, it is the innate ability to harness one's experience that distinguishes an expert from a novice.

Today's optimization algorithms largely overlook this facet of human cognitive ability. Associated AI algorithms therefore lead to prohibitively high sample-complexity in re-evaluating vast spaces of possible solutions. This has caused unsustainability concerns, especially in domains where obtaining data is expensive. The topic of *transfer and multi-task decision-making* has emerged to bridge this gap[269 - 271]. It involves the automated transfer and adaptive reuse of data, learned models or other forms of experiential priors from one task (the *source*) to the next (the *target*) to achieve efficient problem-solving. This ability to build on prior knowledge can greatly reduce computational and data requirements in discovering novel yet high-quality solution prototypes, therefore positively impacting a wide range of applications in science, engineering and manufacturing.

Scientific Background

The literature in this space is broadly categorized into two sub-topics, namely, (a) transfer evolutionary and Bayesian optimization, and (b) evolutionary multitasking and multi-task Bayesian optimization.

The phrase *"transfer evolutionary optimization"*[272] refers to a new class of randomized evolutionary algorithms (EAs) designed to automate machine-machine knowledge transfers in such scenarios. The uniqueness of "sequential" transfer optimization lies in the assumption that tasks occur in a temporally separated manner. Hence, when tackling a target task, computationally encoded knowledge (in the form of data and/or learned models) from one or more previously encountered sources is deemed available for transfer and reuse.

EAs are inspired by natural evolution, a powerful problem-solver that has remarkably brought to life Homo Sapiens from a primordial soup of elementary atoms and molecules. In AI and artificial life, an EA serves as an *in silico* problem-solving equivalent – commonly used for search and optimization – inspired by fundamental principles of natural selection or "survival of the fittest"[273]. This salient feature lends itself well to the transfer optimization paradigm. The general strategy is to seed evolutionary search processes with experiential priors for the optimum. If the priors are useful, the algorithm preserves and refines them over subsequent iterations. If the priors do not contribute positively to the solving of the target task, then the in-built selection pressure sieves out all irrelevant information.

The transfer of priors has been investigated for automated evolution of computer programs themselves[274]. Lately, transfer neuroevolutionary algorithms have been proposed for evolving neural network policies for continuous control tasks, showcasing significant speedups in convergence to near-optimal solutions – and even the discovery of high-quality solutions that couldn't be found within reasonable timescales by conventional algorithms[275]. The ECOLE (Experience-based Computation: Learning to Optimize) programme, a consortium of universities and companies under the European Union's Horizon 2020 initiative, has contributed to solving engineering problems and dynamic multi-criteria optimization tasks[276].

Despite these successes, the applicability of EAs is usually limited to domains where there are no tight caps on the affordable number of sample evaluations; typically, populations of candidate solutions are evolved over several iterations, making thousands of evaluation calls. Hence, under extremely tight resource budgets – e.g., in the order of a hundred or fewer data samples – evolution gives way to a different class of (probabilistic) surrogate-assisted Bayesian optimization (BO) algorithms[277].

Transfer Bayesian optimization follows the modus operandi of augmenting the generalization ability of the probabilistic surrogate model (typically a Gaussian process approximation of the true objective function to be optimized) by combining source and target data and/or models. The terminology meta-BO (short for meta-learning BO) has been used in related contexts, with proven performance guarantees recently made available in the literature[278]. Effective forms of human-machine transfer of priors in BO have also been proposed[279]. Initial works in this area were based on the simplifying assumption that similar source and target problem instances possessed similar looking objective functions[280, 281]. Relaxing this condition, adaptive transfer BO (with multi-program surrogate-assistance) was shown to achieve more than 30% saving in computational cost in designing advanced composite materials in manufacturing processes[282]. Likewise, high degrees of cost savings were reported in the automated design and hyperparameter tuning of modern ML algorithms. Similar use-cases were also unveiled with *Google Vizier*[283] – a Google-internal service for black-box optimization – equipped with a stack of Gaussian process (GP) surrogates drawn from multiple tasks for transfer.

Evolutionary Multitasking and Multi-Task Bayesian Optimization: The field of multitask optimization – extending randomized EAs or BO under the unique assumption that multiple tasks are tackled "concurrently" (in a single algorithm) with periodic exchange of knowledge between them – has received pioneering contributions from Singapore[284, 285]. Although still young, research on evolutionary multitasking is at the core of an expanding corpus of literature on natural computing[286] and AI[287, 288]. A variety of algorithmic realizations have been proposed, confronting questions on what, how, and when to transfer in the unique context of multi-task optimization[289]. To date, the most widely referred among them is the family of multifactorial evolutionary algorithms (MFEAs) that produces the ability to multi-task by simulating the transmission of complex developmental traits to offspring through interactions of genetics and cultural factors[284]. In the alternative BO literature, the associated area of multi-fidelity BO has received significant attention at local institutions[290].

Recommendations

Transfer and multi-tasking are deemed essential to sustainable AI, especially in search and optimization settings where obtaining solution evaluation data is resource-intensive. Examples include AI for boosting scientific discoveries, and searching for creative engineering solutions in vast spaces of possible solution prototypes. However, technical challenges remain in reliably transitioning today's techniques to such real-world use-cases, calling for sustained research effort. These challenges can be grouped into two sub-topics: *Knowledge transfer across heterogeneous problems*, and *Scalability of transfer and multitask optimization algorithms.* When embedding knowledge transfer mechanisms into optimization algorithms, the factors to consider include maximizing beneficial (positive) transfers and minimizing harmful (negative) transfers. These factors can sometimes be mutually conflicting: Solving the tasks in isolation can trivially minimize negative transfers, but this would obviously deactivate all positive transfers as well. Defining *what, how, and when* to transfer to achieve an optimal balance between the two factors is thus highly non-trivial. This challenge escalates in settings with *heterogeneous* tasks whose apparent differences in features conceal useful hidden knowledge. *Scalability* is yet another challenge in today's digital world where computing services are often faced with an explosion of tasks, each with specialized needs. Transfer and

multi-task optimization algorithms therefore grapple with simultaneously satisfying attributes of scalability against a growing number of tasks, and online learning agility against a sparsity of sources relevant to the target[291]. Satisfying these attributes shall facilitate sustainable deployment of algorithms to scenarios with big task-instances, augmenting service level and throughput of computational problem-solving.

Another profound use-case for transfer and multitasking lies in auto-configuring large-scale ML models themselves. For modern deep learning systems, size is power. Massive neural networks trained on broad data are at the forefront of AI. These have come to be regarded as *foundation models* by some[292], enabling fine-tuning for various downstream tasks. Astonishingly however, a life-cycle assessment of such models found that the process of configuring and training them can emit more than 626,000 pounds of planet-warming carbon dioxide[293] – which is nearly five times the lifetime emissions of an average car. In light of the above, transfer and multitask optimization algorithms are expected to play a central role towards low-cost composition of models that uphold environmental sustainability in the years to come. This could be achieved by jointly neuroevolving ecosystems of compact models – e.g., by the evolutionary compression of foundation models to specialize collectively to different task settings, objectives, constraints or user intentions – leveraging the principles of multitasking in just a single algorithmic pass.

EXPLAINABLE AI AND UNCERTAINTY AWARENESS

EXplainable AI (XAI) is a sub-field of AI which emphasizes that the actions, behaviour and solutions provided by an AI system should be understood by humans. This is in contrast with "black-box" models where even the designers of AI systems or architectures cannot explain how they arrived at specific decisions. This topic is particularly prevalent among deep learning AI models as well as hybrid models integrating neural and symbolic AI. Explainability is essential for users to effectively understand, trust, and manage powerful AI applications[294 - 296]. This needs to be in place if human stakeholders are to allow AI to exert more and more influence in their day-to-day lives, including in life-and-death situations such as medical diagnostics and autonomous vehicles.

Research and breakthroughs in XAI approaches could bring enormous benefits including facilitating transferability of machine learning models, enhancing the informativeness of recommendations, increasing users' confidence, increasing fairness of decisions made by AI models, allowing better accessibility to non-technical users of AI, enabling better interactivity between AI and users, and enhancing privacy awareness of users and regulatory entities[297].

Scientific Background

Today's XAI approaches and techniques can be categorized based on their a) Applicability to transparent vs opaque models; b) Stage of explainability (pre-modelling explainability, in-model explainability, and post-hoc explainability); c) Applicability to specific AI models, or model-agnosticity. Figure 5 below is a visualization of the aforesaid categorization. Opaque AI models are typically explained by *post-hoc explanations*, whereas transparent AI models are typically explained during the *pre-modelling and modelling explanations*.

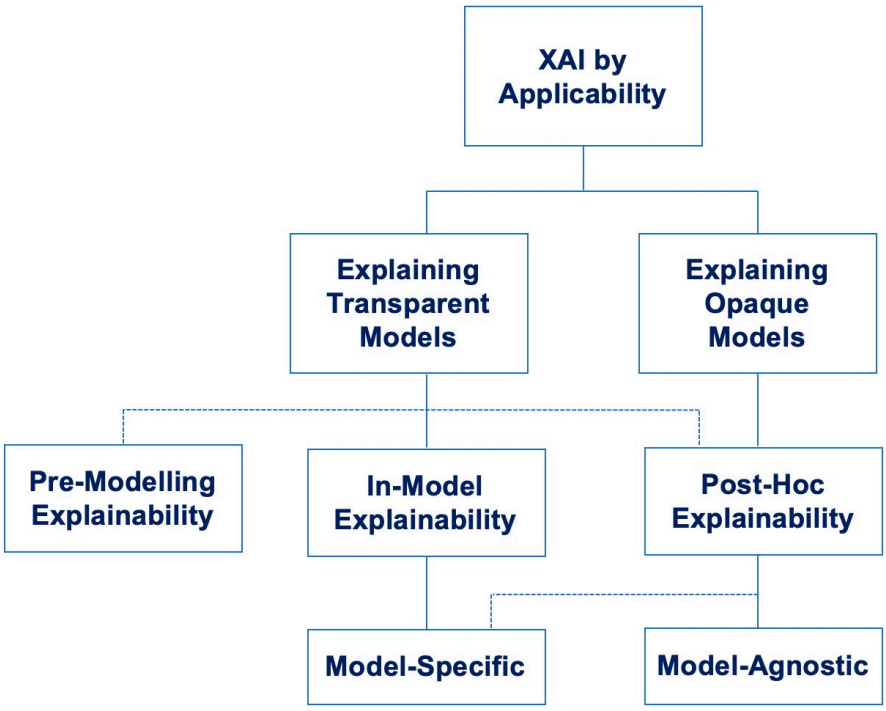

Figure 5: Categorization of today's XAI approaches and techniques

Model-specific approaches are typically designed to work with respect to the intrinsic architecture and features of the target models[90]. This makes them suitable to explain

transparent models. On the other hand, *model-agnostic approaches* are designed to be flexible and do not depend on the intrinsic structure and architecture of a model[295, 298]. They solely relate the inputs to the outputs, making them suitable for explaining opaque models such as deep learning, random fields and SVM.

Besides presenting the outputs of the AI models such as decisions and recommendations in a naturally understandable form, the XAI interface also answers questions such as “Why is the solution chosen?”, “What is the reasoning process?”, “What are the factors contributing to the decision?”, “When is the solution valid?”, “How certain is this decision?”, “How much can I trust this decision?” and “What if the values of some inputs are changes?

Uncertainty-Aware Learning is another emerging field. Humans deal with uncertainties all the time in fields ranging from investment to medical diagnostics to sports and weather forecasting[299]. However, ML and DL are found to be unable to handle uncertainty when the training data and evaluation data are mismatched, which means they cannot be reliably applied without a huge amount of training data[300]. To help them achieve efficient and explainable learning and decision-making without a huge amount of training data, it is imperative to improve the models’ ability to handle uncertainty.

Recommendations

There is a fundamental trade-off to be made between explainability and accuracy of AI systems. Symbolic AI and transparent AI models are easier to understand, but do not produce state-of-the-art performance. The high performing AI systems leading the pack are typically based on deep learning opaque models, involving deep network architecture with hundreds of billions of parameters. There is an ongoing race to continue building AI models with increasing size and complexity. This will continue to pose big challenges to explain such increasingly complex AI models.

A general lack of agreement on the vocabulary and definitions surrounding the field of XAI is another limitation. More importantly, the multiplexity of the approaches and methods has led to a lack of agreement, consistent understanding and effort towards establishing objective metrics for evaluation of explanations. A further bottleneck is that most work on XAI today serves the needs of model engineering and development, but do not consider the domain and user-specific factors, making the explanations unsuitable for non-technical users. It is thus critical that XAI should incorporate human consideration into the design so as to enhance AI trustworthiness and usability.

One approach to explainability is to endow an AI system with the ability to learn and operate with symbolic representations. Classical AI has started out by using symbolic representations, but due to the inability to learn those symbolic representations, the symbolic approach fell out of favour. Hence, one major challenge is to devise new methods to learn concepts in symbolic forms. Another issue related to symbolic representation explored in the early years is that the symbols used were not grounded. In other words, the meanings of the symbols were not adequately represented and there was no systematic and principled way to ensure that the AI system truly understood the symbols. Hence a major challenge going forward is to create XAI systems based on grounded symbolic representations and also to ensure that these representations are amenable to learning.

AI systems provide the best value by incorporating domain-specific considerations in the decision-making process. Accordingly, explanations for AI systems in specific domains can take advantage of the vast knowledge available. There have been efforts in Singapore toward domain-specific XAI, most notably in the medical and healthcare domains[122, 124].

CAUSAL INFERENCE

Much of today's state-of-the-art AI models, such as deep learning models in particular, rely heavily on the discovery of the correlation between data in their attempt to learn meaningful patterns or insights from those data. *Correlation* is when two things are true at once, but one does not cause the other. In contrast, causation is about establishing stable, invariant relations where one factor causes another.

For AI models and systems to progress towards explainability and interpretability, AI needs to upgrade from depending mainly on correlation to incorporating *causality* as well. This is so that the performance of AI is less likely to be adversely affected by spurious correlations which lead to inaccurate or false *inferences*. For example, countries with high chocolate consumption also happen to be countries that win a high number of Nobel Prizes, but this is a *correlation*. It would be obviously wrong if an AI were to conclude that high chocolate consumption *causes* the high number of Nobel Prizes, when the underlying common cause is the GDP.

Therefore, AI cannot explain the true cause by only using correlation. If AI's reasoning becomes based on transparent cause-effect relationships and not just opaque black-box probability, the issue of AI using potentially spurious correlations to perform unexplainable decision-making can be mitigated. Beyond statistical learning, reasoning with rule-based, knowledge-based and logical inference is important to realize human-like deep reasoning skills from limited information. Core statistical methods such as Bayesian reasoning could be explored to understand causation instead of using correlation.

More specifically, causality aims to establish stable, invariant, causal relations among data. Once a causation is established, it can be re-used in any domain, any task, any model assembly, without the need for new data collection, re-training, and task re-definition. Drawing an example from physics, once a system has learned the law of gravity (causation between mass and force), it can reuse it in any place (e.g., Earth, the moon or Mars) and in any field (e.g., aerospace or aviation). When it comes to AI systems whose decision-making will depend on concepts far more complex than gravity, causality becomes even more important.

Scientific Background

In ML AI literature, causality based on graphical causal models[301] is raising significant interest in machine learning community in recent years. Some works try to build the foundations of causal theory, including independent causal mechanisms[302 - 304] (the basis of causal representation learning) and causal discovery[305 - 308] which helps to reveal the underlying causalities from observational data. Causal representation learning[309 - 311] tries to learn the variables from a causal graph into representations, which can be seen as the final goal of learning disentangled representations[312, 313]. 3) *Invariance learning*[314] argues that a causal mechanism remains invariant when other mechanisms are subjected to external influences.

Causality has widely inspired several studies in computer vision tasks. These include traditional vision tasks as well as vision-language tasks. Causality also has been applied in many NLP tasks[315, 316]: Some methods use topic models or autoencoders to encode the underlying confounder in text and perform backdoor adjustment to remove the confounding bias and realize causal. Other methods use counterfactual thinking to create a counterfactual generation model to augment the training data[317 - 319] or directly use the counterfactual inference to adjust the causal effect[320].

In recent years, there has been an attempt to introduce the notion of causality in the field of scientific machine learning or physics-informed learning. These include applications in areas

such as particle physics, materials science and climate modelling where the collection of immense observational datasets makes AI techniques particularly attractive[321, 322]. As big data is not available in many real-world scientific applications, there are efforts to incorporate scientific theories, physical governing equations or known characteristics of systems into AI models[323 - 325]. The insertion of such knowledge is expected to improve model performance, robustness and generalizability even under such data-scarce conditions.

Recommendations

To evaluate the results of causal inference, the development of appropriate testbeds and metrics is crucial. There is a current lack of proper standards for evaluating a machine's causality. Causal inference research should focus on developing a family of tasks for causal machines, setting standards for AI's trust, rationality and morality. Furthermore, below are several promising technical topics to achieve progress in the rationalizability of AI systems based on causal inference.

Feature disentanglement is necessary for intervention in an AI model's causal inference. Almost all existing feature disentanglement methods are based on statistical independence, which is an unnecessarily strong assumption. For example, in face recognition, "eye" and "nose" are disentangled parts, however, they are statistically dependent. Relying on such dependence may wrongly result in coarse disentangled parts such as "eye and nose" vs. "sunglass". A standard evaluation metric is also needed for assembly of new AI models by reusing (disentangled) parts trained by others. Therefore, "disentanglement ability" needs to be integrated to define an evaluation metric agnostic to the ground-truth labels.

As opposed to conventional end-to-end Supervised Learning (SL)[326], Self-Supervised Learning (SSL)[327, 328] first learns a generic feature representation (e.g., a network backbone) by training with unsupervised pretext tasks such as the prevailing contrastive objective, and then the above stage-1 feature is expected to serve various stage-2 applications with proper fine-tuning. SSL is important for the creation of foundation models trained on vast quantities of data at scale that show high generalization ability in different tasks. SSL for visual representation[329] is a breakthrough in that it is the first time that "good" visual features can be obtained for free, just like pre-training in the NLP community. It was found that most SSL work only focuses on how much stage-2 performance an SSL feature can improve. Thus, there's a need for more research into important questions involving what features SSL is learning, how can these features be learned, what features can or cannot be learned, what the technical gaps between SSL and SL are, and how SSL can surpass SL.

Invariance is another key area that needs exploration. Intervention is about the pursuit of invariant predictors across domains. Given a task, e.g., classification or regression, there is a need to define the meaning of "invariance" in that context. Existing methods have yet to create a formal definition of "invariance", as well as an algorithmic guarantee. A theory of "invariance" vs "equivariance" is needed to overcome the limitations of existing statistical learning AI approaches by (a) identifying the objective of "invariant predictors", and (b) encoding "data dynamics" – transformation – into the formulation of learning tasks.

AI should not just make predictions, but also advise users on best courses of action. To do this, AI needs to imagine unseen scenarios as counterfactuals such as "but for" and "what if". For example, an AI system used by a bank to assess loan applicants is limited if it just predicts that an application will be approved or rejected. It is more useful if it advises the applicant what can improve their chances of getting their application approved, and also provides the bank and regulator with recommendations from different perspectives.

Another area where progress is needed is the ability for ML to handle dynamic data. Current ML technology relies heavily on statistically independent data. But real-world data is often time-dependent, and ML needs to become better at processing this.

At a more macro level, understanding the mathematical foundation of intelligence would be a challenge worth pursuing. Since the success of deep learning has not been fully understood yet, elucidation of its mechanism is essential for future development, including but not limited to: statistical understanding of deep-layered information processing, the capability of solving non-convex optimization problems with stochastic gradient descent, and better generalization in higher-dimensional cases.

COGNITION AND NEUROSCIENCE

Cognition is the study of the human mind and brain, while Neuroscience is the study of the nervous system and its functions. In recent years, these fields have been gaining attention in AI research as they are increasingly deemed relevant to creating better AI especially towards the notion of Artificial General Intelligence (AGI)[330]. The human brain is the main existing evidence that this kind of generalized intelligence is even possible. As such, scrutinizing the inner workings of human brains through cognitive science and neuroscience is one of the most promising approaches to pursue more powerful AI including AGI.

Imbuing AI with human-level cognitive capabilities and intelligence increases the possibility of AI technologies yielding greater economic and societal benefits. AI research that incorporated cognitive science and neuroscience aspects have already helped create two of the most dramatic advancements in AI with rule-based systems articulated by experts and data-driven approaches with deep learning. The benefits can be enormous and bidirectional – studying human intelligence facilitates the creation of biologically-plausible AI, while studying AI helps one better understand and possibly augment human intelligence and potential.

For instance, AI approaches such as learning paradigms, neural network architectures and symbolic processing have also helped the study of cognitive capabilities in biology, including human intelligence. A better knowledge of human intelligence stands to increase understanding of neuropsychiatric disorders and mental health. This can potentially lead to early detection of at-risk individuals, improve diagnosis, predict disease progression and develop treatments. Moreover, the development of better BCI technologies can also help restore certain cognitive or motor function to patients, or even augment the intelligence of healthy individuals.

Scientific Background

A major approach to understand the human mind is by studying *behaviour* of individuals and groups – giving them behavioural tasks and measuring their performance (e.g., accuracy, reaction times etc) to gain insights into the mechanisms underlying different cognitive processes, including perception, memory, attention, decision-making, language etc. In parallel, systems neuroscientists seek to elucidate *biological circuitries* underlying these cognitive processes. A broad array of techniques (e.g., electrophysiology, imaging, etc) are employed to decipher neural circuitries across multiple spatiotemporal scales, from single neurons to large-scale networks.

More recently, there has been significant development in *neural technologies*. Efforts can be divided into two types: neuromorphic computing and neurotechnologies. *Neuromorphic computing* is the development of silicon hardware to emulate neural systems. On the other hand, neurotechnology involves the development of devices to interface with the nervous system. Recent technologies, such as neuropixels, allow the simultaneous recordings of hundreds of neurons, which could potentially be enhanced to allow the recording of thousands of neurons across many regions[331].

There are two possible ways of achieving biologically-plausible AI: (a) Emulating human behaviour, and (b) Emulating human brain circuitry. AI models that emulate human behaviour seek to mimic manifestations of human cognitive function – such as Neural Turing Machines[332]. The concept of attention (the cognitive process of selectively focusing on some information) has influenced an entire subfield of deep learning architectures[333]. On the other hand, AI models that emulate human brain circuitry refer to those that seek to mimic certain

aspects of brain circuitry that support human cognitive functions. For example, convolutional neural networks (CNN) mimic the hierarchical structure of visual processing in human brains. Spiking neural networks (SNN) are artificial neural networks (ANNs) that are considered more biologically realistic than other ANNs[334] and are known to achieve higher computational efficiency[335].

AI has improved understanding of human intelligence in at least two ways: (A) Using AI models as models of brain function and behaviour, and (B) Using AI to make sense of brain data. There is an influential perspective that suggests that to understand biological intelligence, it is necessary to figure out the underlying objective function, learning rules and biological architecture that subserve animal cognition[336]. One approach is to correlate the activity of ANNs with actual brain activity in humans and non-human primates. There are also efforts to compare the mistakes of ANN with animal behaviour in order to gain insights into diverse cognitive functions.

Increased understanding of biological intelligence, better sensors for recording brain signals and advanced AI techniques have led to advances in BCI. As an example, the brain signals of participants silently miming sentences have been successfully decoded into synthesized speech that human listeners can make sense of[337]. Brain signals of participants imagining handwriting have also been successfully decoded to text at a rate comparable to smartphone typing speeds[338]. Artificial DNNs have been used to predict visual stimuli that maximally control the activity of certain brain regions[339, 340].

Recommendations

Many aspects of current AI remain quite different from humans. For example, human intelligence is more robust in partially observable or perturbed environments, changes of contextual information and distributional shifts. These might be due to missing top down cognitive processes (e.g., executive control, world model, beliefs, motivations, goals, intentions). Moreover, current AI models are also quite domain-and-task-specific (e.g., distinct models for computer vision and natural language understanding). Although there are increasing attempts to build multimodal and multitask AI (e.g., text-image models), these models are still “passive observers”. This is in contrast with human intelligence, which emerges from the interaction of agents with their environments and other agents across multi-modalities as a result of sensory-motor activities.

State-of-the-art AI remains biologically non-plausible. It is currently unknown what aspects of biological brain circuitry are quirks of evolution, rather than key components of intelligence. So far, efforts to include more biologically-plausible components have led to AI models that are more similar to brain data, but task performance is at best similar to (but not better than) state-of-the-art AI models. However, it is unclear whether the added biological components are not necessary for intelligence or whether the implementation is lacking additional key elements. After all, feedforward deep neural networks languished for decades not because the idea was bad, but because the required components (data and computing power) were not ready. Therefore, given that the human brain is still one of the few sources of general intelligence, such efforts should be continued.

There have been significant recent efforts to build datasets that can be used to train AI models with accompanying brain and behavioural data. This facilitates the evolution of more biologically realistic AI models. These efforts are in the right direction, but are mostly restricted to processing in the ventral visual stream. Testbeds in other areas of brain function (e.g., language, reasoning) will be important.

Using AI to analyze brain data involves processing followed by analysis. ML has been increasingly important for both these stages. However, while deep neural networks have been increasingly successful for processing, classical ML remains dominant for analysis. One reason is that in pre-processing (e.g., segmentation), goals and "ground truth" are often well-defined. However, in analysis, a key component is the interpretation of the results, and modern AI models remain harder to interpret than simpler classical ML models. For BCI systems, developing stable systems that can be continuously used for long periods without model retraining remains a challenge. Non-stationarity in brain signals occurs due to tissue scarring, micromotion of electrodes and changes in environmental noise. Adaptive AI methods and new advancements in signal processing could address some of these issues. There has been increasing interest in the intersection of AI and cognition/neuroscience. A notable recent funding initiative on the global stage was from Priscilla Chan and Mark Zuckerberg to Harvard to create the new "Kempner Institute for the Study of Natural and Artificial Intelligence, which is a new University-wide programme seeking fundamental principles that underlie both human and machine intelligence."

In Singapore, funding opportunities for such research are mainly through health-related funds (e.g., Singapore National Medical Research Council) and MOE Tiers 1-3. At present, the vast majority of cognition/neuroscience projects are focused on healthcare applications such as mental health, but not immediately on how AI can be involved in the studies. The research challenges and opportunities discussed previously could serve as potential areas for more funding and government support. The study of cognition and neuroscience in AI is a highly interdisciplinary one, requiring expert knowledge in both areas. While there are promising local researchers in Singapore, they are spread out across multiple institutes, which poses operational difficulties for these researchers to collaborate in a truly interdisciplinary fashion.

Curated dataset (e.g., ImageNet) have been critical to rapid development of AI, but are much less common in areas involving cognition and neuroscience. Developing *open AI-ready testbeds* for this topic will be equally important. These should have standardized inputs, outputs and data. Further, they should be available under open licenses that allow researchers to seamlessly inspect, reuse, standardize, and extend the open-source data repositories, which also reduces redundancy of efforts and cost.

A barrier to the development of such efforts is that collecting and sharing of human/animal data necessarily involves navigating significant regulations involving institutional review boards (IRB), institutional animal care and use committees (IACUC) and research collaborative agreements (RCA). Given the lack of experience of most computer scientists or engineers in dealing with IRB/IACUC, this is a major barrier for collaborations between AI researchers and neuroscientists. *Streamlining of IRB, IACUC and RCA procedures for the collection and sharing of de-identified data is recommended*.

Even though current AI technologies are not near human-level intelligence, they are shown to have profound impacts in multiple industries, such as healthcare, manufacturing, transportation, education and others. Following the pervasiveness of AI in everyday life, it is essential to consider the topic of HAS in the next section, which looks at humans working together with AI towards greater goals of economic and societal benefits.

HUMAN-AI SYNERGY

AI has exceeded human performance in various recognition tasks, but its brittleness (caused by the lack of robustness to contextual changes for instance) often hampers its wider adoption in real world-tasks, especially when there are safety implications and user acceptance considerations. This is why collaborative human-AI workflows become more important to achieve large-scale deployment and adoption of AI technologies. This gives rise to the topic of *Human-AI Synergy (HAS)*, which involves humans and AI working together to achieve a complementary combination of human intelligence and artificial intelligence, so that this collective intelligence significantly expands the capabilities of both machines and humans, and achieves shared goals.

Human-AI Synergy is envisioned to herald a new era in the development of AI characterized by trust and understanding between humans and AI. AI working with humans can deliver improved quality and performance while reducing human workload and stress[341] in areas such as medical diagnosis[342], copywriting[343] and customer service[344]. Compared to autonomous AI which is often perceived as opaque "black-boxes" lacking human understanding, collaborative AI that understands and works with humans would be much more likely to receive greater acceptance. The idea of AI working with people rather than replacing them in jobs would also increase society's confidence in using these technologies at work and in their daily lives.

Figure 6 below from Deloitte[341] shows the improved outputs and capacity that can be achieved when humans and AI work together. At the first stage, substitution, new outputs result in reduced costs and improved efficiency. At the second stage, augmentation, a greater degree of transformation delivers greater value and expanded opportunities in addition to reduced costs and improved efficiency. At the third stage, collaboration, an even greater degree of transformation lets the work and outputs take on more meaning for workers and customers, as well as delivering greater gains in costs, efficiency and value[345].

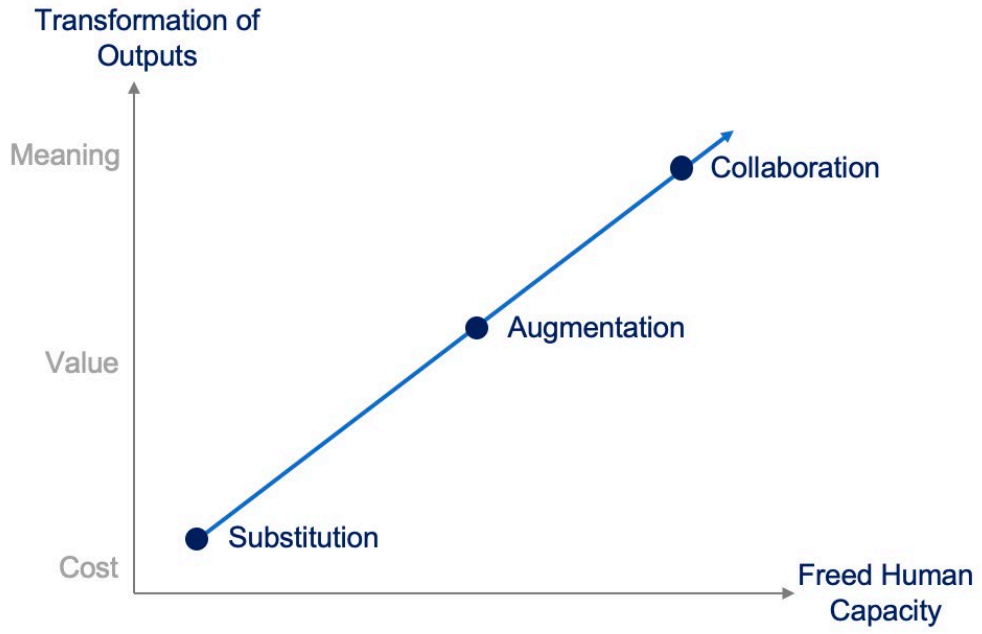

Figure 6: Stages of human-AI synergy and their levels of output and capacity

Scientific Background

HAS is an emerging multidisciplinary field[346]. Current research (other than AI itself) is scattered across different communities. On one hand, the study of humans involves the fields of cognitive science, psychology, linguistics, organizational behaviour, etc[347]. On the other hand, the development of Human-AI systems themselves is currently carried out in the areas of human-computer interaction[348], human factors engineering[349], human-robot interaction[350], etc.

Examples of human and social science, and human factors topics are: a) Implicit understanding of humans (including theory of mind and commonsense knowledge about human abilities) for non-verbal interaction; b) Rapport building for establishing trust, pre-empting human needs etc., to ensure fluent co-ordination; and c) Negotiation, persuasion, etc. (including affective understanding, motivation theory, empathetic communication, and theory of negotiation) for smoother social interaction.

Examples of interaction topics are: a) Goal alignment and constraint identification (including task and goal representation and tracking, and understanding of human abilities); b) Sensemaking (such as task and situation understanding); c) Interface design (including models and theories of human-machine interaction use and conceptual frameworks for the design of interfaces); and d) Planning and control (such as (re)allocating tasks/subtasks between humans and AI).

It is worth noting that many of these topics have an underlying goal of building trust and understanding between humans and AI[351]. Trust and understanding are also necessary for AI to become more usable (e.g., human-centric explainability, natural interaction with commonsense) and useful (e.g., for AI to be applied to a wider range of real-world tasks that include complex social interactions). Human understanding also extends beyond individuals to social understanding of groups and social contexts. Related work can be found in social robotics, social computing[352] and the newly-proposed field of socio-neuro AI[353].

Some of the recent research trends related to HAS include Human-Robot Collaboration[354, 355], Human State Sensing, Human-AI Multiplayer Competitive/Collaborative Games and Multi-Agent Reinforcement Learning, Humans Adapting to Chatbots and Software Agents, Human-Agent Collaboration[356], and AI Assistants.

Recommendations

As previously highlighted, HAS research covers multiple disciplines and fields involving human and machine cognition. Apart from engineering efforts required to integrate across different modalities and intelligences (visual, auditory, spatial, etc.), there are additional aspects of how collaboration arises from human and machine cognition that are still not well understood and require systematic scientific investigation and research.

The research challenges in HAS can be organized into three categories, namely, Explicit Challenges, Underlying Challenges, and Practical Challenges. Explicit Challenges relate to functional capabilities needed for effective human and AI collaboration. Underlying Challenges include lack of commonsense and context (where background knowledge about the world/social situations that is shared by humans is lacking in AI), insufficient progress in embodiment (AI with a physical body) and embodied sensing (real-world perception), the lack of shared mental models, and heterogeneous entities at the individual and group level. Practical challenges include safety concerns, computational resource constraints, and the need for remote, asynchronous and non-stepwise collaboration capabilities.

For Singapore, HAS allows AI to be deployed to address critical national challenges such as its ageing population and labor shortage. Although AI has made significant advances in the past decade, human-like AI capable of handling complex real-world challenges has not been realized. This is why harnessing AI capabilities in synergy with humans to address complex issues is important. Furthermore, Singapore's educated workforce is a key competitive advantage that can help the country increase its innovation speed and capability through HAS. Its advantages include an existing pool of researchers in various disciplines, supportive governmental measures, a pragmatic and rigorous approach to investment in research and capability development, and a trustworthy brand name and reputation as an attractive location for expats which helps in talent attraction and retention.

LEARNING AND ARTIFICIAL EVOLUTION

The human brain is a prime specimen of the power of evolutionary natural selection. It is among the most potent enablers of intelligence and lifelong learning in the natural world, serving as a primary inspiration to the field of AI as a whole. From a computer science perspective, biological evolution can be thought of as an *outer-loop* optimization algorithm that, over many generations/iterations, has shaped increasingly *fit* brains. The *Baldwin effect*[357], in evolutionary biology, describes how this notion of fitness may stem not only from capabilities derived from the brain's inherent structure – such as in precocial species whose young already possess certain skills from the moment of birth – but also its capacity to engender *inner loop learning* during an organism's lifetime. Drawing parallels to the field of deep learning, a brain's structure loosely translates to the architecture of a deep neural network model, whereas inner loop learning typically takes some form of backpropagation via stochastic gradient descent (SGD) for tuning the network's parameters under observational data.

It is this interaction between randomized EAs and gradient-based optimization that lies at the heart of the proposed synergy between learning and artificial evolution. Just as (slow) biological evolution has been responsible for shaping intelligence in real life, fast in silico *neuroevolutionary algorithms* (i.e., EAs for training deep neural networks) are expected to play a role in shaping the future of AI in artificial life.

Scientific Background

The growing literature synergizing learning and evolution in AI can be categorized into three sub-topics, namely, (a) artificial evolution alongside gradient-based learning, (b) evolutionary reinforcement learning, and (c) evolution of ML subsystems.

The success of DL is driven by the observation that under sufficient training data and processing power, simple backpropagation which applies stochastic gradient descent (SGD) to minimize a differentiable error/loss function is effective for training neural networks with billions of parameters. However, SGD is no silver bullet. It is susceptible to stumbling blocks such as local optima, saddle points, and error plateaus[358]. Such shortcomings illuminate the scope of gradient-free global optimization for training neural networks. Neuroevolution is a dominant example of this algorithmic class[359].

Beyond attempts to arrive at substitutes to SGD, a concurrent research track proposes to *synergize global evolutionary search alongside local gradient signals* – thus inheriting the advantages of both approaches. For example, researchers from Google DeepMind revisited the Baldwin effect through the lens of neuroevolution, showing its ability to shape initial parameters of deep learning algorithms to enable rapid (few-shot) adaptation to empirically challenging tasks via gradient-based finetuning[360]. Likewise, researchers from the IBM T. J. Watson Centre proposed to hybridize SGD with gradient-free EAs as complementary algorithms within a single framework for optimization[361]; the motivation is to meld SGD's ability to exploit curvature information of error functions, with an EA's capacity to explore complex function landscapes.

Evolutionary Reinforcement Learning:

Reinforcement learning (RL) forms a class of problems where AI agents are trained to act in dynamic environments. An agent's behaviour is governed by a policy function (the agent's brain) which computes the action to be taken in each situation. In deep RL, these policy

functions take the form of deep neural network models trained with the objective of maximizing cumulative rewards. OpenAI's recent demonstration of the effectiveness of evolution strategies for deep RL – attaining competitive results on Atari games after just one hour of training in a distributed setting[362] – has given the field a fillip.

Efficient use of sampled data could also be achieved via iterative procedures that optimize surrogate objective functions, providing monotonic improvement guarantees[363]. Randomized evolutionary operators following principles of imitation learning have been designed to preserve the hierarchical relationships of neural network parameters, lessening the danger of catastrophic performance drops[364]. The hybridization of reward-guided evolution with explicit search for novelty[365] is yet another approach where RL agents are encouraged to exhibit different behaviours, hence reducing the danger of being stuck indefinitely (and hence wastefully) in local optima of deceptive reward functions.

Evolution of ML Subsystems:

In evolutionary biology, the Baldwin effect describes how natural selection is affected by an organism's ability to learn during its lifetime, lending selective advantage to those traits that enhance learning[366]. In the context of AI, this theory motivates a range of techniques for outer-loop optimization of ML subsystems[367] – also referred to as AutoML. Applications include hyperparameter optimizations of ML models[368], the search for optimized neural network architectures to be trained via SGD[369], evolutionary compression of massive pre-trained neural networks[370, 371], and even the discovery of complete learning algorithms from scratch using only basic mathematical operations as building blocks[372]. A related research track explores the evolution of embodied AI, where morphologies of artificial agents (e.g., soft robots) are evolved in an outer optimization loop to facilitate the learning of tasks (e.g., via an inner RL loop) in challenging environments[289].

Recommendations

As alluded to earlier, the potential of synergizing evolutionary search alongside local gradient signals – thus producing AI that inherits the best of both worlds – has garnered global attention. However, the construction of such integrated frameworks poses fundamental challenges in specifying precise interaction mechanisms between gradient-based and gradient-free learning. Notably, the study of such interaction mechanisms forms an integral part of *memetic computation* with EAs – a subject with a rich history and pioneering work ongoing at Singaporean institutes of research and higher learning[373, 374].

Given the increasing interest towards creation of general-purpose AI systems, there is also a shift from "internet AI" that deals with learning from datasets of images, videos and text curated from the internet, towards embodied AI which enables artificial agents to learn through interactions with their surrounding environments[375]. Embodied intelligence is the belief that true intelligence emerges through the interplay of an agent with physical environments characterized by multimodal data streams (e.g., simultaneous vision and audio), mediated by the constraints of its body (i.e., its morphology), sensory and motor system, and brain. Notably, today's AI models and applications are mostly limited to processing data of only a single modality at a time. In contrast, actualizing the envisioned future of multimodal embodied AI systems could have unprecedented impact on strategic domains of Singapore's RIE2025 plan – including advancing robotics technologies to transform the built environment in a sustainable manner or enable innovation and enterprise talents to leverage robotics for various industry needs. The synergy between learning and evolution to embody AI into adaptive robots is expected to play a central role in taking this vision to reality. Just as natural evolution has

shaped a variety of specialized organisms occupying different environmental niches, artificial evolution may also be capable of generating diverse structures for body and brain, optimized to carry out different tasks jointly or independently.

Despite the immense impact potential of this overarching idea in the real-world, as of today, embodied AI research has largely been focused on learning only in virtual environments. This trend has led to significant advances in simulation engines that aim to faithfully replicate the physical world. These *simulated worlds* serve as virtual testbeds to collect task-based datasets[376] and to train and test AI frameworks, alleviating the threat of deleterious outcomes in unknown physical environments. Several simulators have therefore been developed over the past four years[377 - 380], providing realistic representations of the world. Most of these simulators minimally comprise a physics engine, Python application programming interface (API), and an artificial agent that can be controlled or manipulated within the environment. While this fosters immediate technical advancement, open questions have however arisen in terms of *closing the reality gap*, for safely transferring evolved agents from simulation into the real world[381].

AI INFRASTRUCTURE

A common theme that has been observed among the major AI R&D initiatives worldwide is the substantial global investments in High-Performance Computing (HPC). For instance, Meta is currently building the world's fastest AI supercomputer, made up of 16,000 GPUs with a capacity of 5 exaFLOPS. Asia's largest AI data centre has opened in Shanghai in 2022 with a compute capacity of 3.74 exaFLOPS.

The *National Supercomputing Centre (NSCC)*[382 - 384] is home to Singapore's national petascale supercomputers known as *ASPIRE 1 and ASPIRE 2* (Advanced Supercomputer for Petascale Innovation Research & Enterprise (*ASPIRE*)). ASPIRE 1 was benchmarked at 1 PetaFLOPS (PFLOPS) and provides 275 million core hours of usage per year. It contains a 13-Petabyte high performance data storage system and is interconnected to other facilities via advanced networks: locally (40-100G), regionally (10-100G) and globally (10-100G) through partnership with Singapore Advanced Research and Education Network (SingAREN). ASPIRE 2 is expected to provide up to 10 PFLOPS of computing capacity and is 8 times more powerful than ASPIRE 1. While ASPIRE 2A is designed for HPC applications, its CPUs and 352 x Nvidia A100 GPU cards can also run high throughput computing (HTC) and AI applications.

These two supercomputers are crucial to Singapore's future supercomputing resources which will support research in areas such as climate change, biomedical research and smart nation activities. The NSCC plays an important role in facilitating *sustainable AI infrastructure*. As a centralized HPC resource, NSCC's ability to optimize its entire operation and achieve energy efficiencies can be translated to carbon footprint reduction for its ecosystem of users. It could also help to reduce the redundancy of each individual user's HPC cluster hence reducing the corresponding carbon footprint.

Scientific Background

HPC-based computer simulation is often referred to as the "third pillar" of scientific discovery, complementing traditional theory and experimentation. The usage of HPC resources has spread from its established strongholds in the physical sciences to social sciences and the humanities, enabling significant impacts such as saving lives and property by predicting severe storms, reducing the time-to-market, and increasing safety and reliability in the automotive and aerospace industries. For these reasons, leading nations are investing substantially in supercomputing[385] and emerging disruptive infrastructures such as AI accelerators and Quantum technologies. In contrast, Singapore's HPC investments have remained lower than those of other nations over the last decade – which urges the need to consider both the benefits of increased investments and the risks involved by not doing so.

High-Performance Computing:

To democratize the use of HPC and AI, investments in software and applications need to dovetail with hardware. Some of the recommended focus areas would include areas that Singapore is already strong such as Smart Cities and Finance. While Singapore has taken the lead in Smart Cities, the same rate of progress has not been observed in HPC applications to support such initiatives including vehicle traffic management, smart buildings and power grids, for instance. This is a crucial aspect to reduce Singapore's potential reliance on foreign capabilities to support HPC-based research.

Singapore's infrastructure investments need to be dramatically increased to maintain its competitiveness in AI R&D. As an example, Hyperion[385] has recommended that Singapore should consider growing its largest 2 to 3 supercomputer investments by at least 30% to 40% a year for each of the next 3 years. Moreover, the need for greater AI infrastructure investments is also motivated by the increasing interdependency among government agencies, IHLs and industry partners – which necessitates a national infrastructure such as NSCC to allow joint collaborations and access to infrastructure. This will also provide the ability to aggregate resource demands and to negotiate with global vendors for the latest technology in the market (e.g., Nvidia's GPU technology). A national shared HPC system will also address the need for a sustainable AI infrastructure as it will reduce the need to build multiple clusters that cause higher carbon footprint, while achieving the best ROI for Singapore. To further empower sustainability, it would be good to build on the ASPIRE 2A supercomputer's significant efforts in this space such as its 'warm water cooling system' that captures 60-80% of the servers' heat and reduces data centre cooling costs by over 50%. The upgrading of national supercomputers, Cloud infrastructure and availing the latest HPC applications are also critical to accelerate AI R&D in key domains such as advanced manufacturing, healthcare, built environment and fintech.

Global partnerships can be further encouraged. An example of an ongoing global partnership is the NSCC's collaboration with Japan's Research Organization for Information Science and Technology (RIST) which enabled Singapore researchers to regularly access supercomputing resources from the world's most powerful supercomputer, Japan's Fugaku system. This contributed to the development of HPC in both countries. Furthermore, establishing partnerships with regional and global HPC centres can help broaden the experience of the Singapore HPC community by getting access to HPC technologies which may not be available in Singapore. Similar partnerships should continue to be established.

Quantum Artificial Intelligence:

A potentially disruptive and useful computing technology which Singapore should pay attention to is the emerging area of *Quantum machine learning (QML)*[386] – which synergizes theories of quantum mechanics with ML methods. QML extends the pool of quantum computing hardware for machine learning by utilizing an entirely different class of computing device known as the *quantum computer* – which uses properties of quantum mechanics such as superposition, interference and entanglement for processing information[387]. QML is mainly motivated by the huge volume of data required to ML systems. With the size of datasets constantly increasing and Moore's Law tapering off, current computational tools may no longer be sufficient for the future[388]. Hence quantum computing (QC) provides a promising alternative for the scalability of AI research with the growing abundance of knowledge and data available. QC's ability to interpret data far more effectively than ordinary computers results in a shorter learning curve for AI robots[389]. QC also promises to render today's data encryption methods obsolete through new encryption technologies such as the Quantum Key Distribution (QKD)[389].

There are four ways to combine QC and ML, based on whether the data is generated from a quantum (Q) or classical (C) system, and if the information processing device is quantum (Q) or classical (C). *Hybrid Classical-Quantum (CQ) ML* involves converting classical data into quantum data and going through the 3 phases of encoding, processing and measuring. Data is obtained from observations from classical systems such as text, images and time series data, which are input into a quantum computer for analysis. *Purely Quantum (QQ) ML* is similar to CQ and considers quantum data processed by a quantum computer. The data could be obtained from a measurement of a quantum system in a physical experiment and transferring the measurements into a separate quantum computer. Alternatively, the dataset could comprise of quantum states[390] such as the final quantum states of a quantum dynamical

simulation[391]. *Quantum-Inspired (CC) ML* uses the conventional approach to ML, where the classical data is being processed by classical computers. However, unlike conventional ML, CC ML borrows methods from quantum information theory, for example by utilizing tensor networks. In the past, CC ML catalogued a body of literature with different degrees of quantum mechanical rigour. *ML-aided (QC) Quantum Computing* looks at how classical ML can aid quantum computing. It largely employs classical ML techniques to analyze quantum states[392]. Some of its methods include using artificial neural networks to represent variational quantum states[393], and using fully connected and convolutional neural networks to identify phases, phase transitions and non-trivial crossovers between topological phases[394].

Recommendations

Continued research and innovation in AI require ongoing investments into AI infrastructure including a comprehensive data strategy, more advanced supercomputers and emerging technologies, and training and development of talent, software and applications. Powerful and advanced AI infrastructure is pivotal to the talent-attraction and retention ability of Singapore and therefore its continued leadership position in AI. As such, investments into computing infrastructure for AI research can enable Singapore to strengthen its position as the Global-Asia node of technology, innovation and enterprise, as well as to reap economic gains from investments into its current research infrastructure.

Data is fundamentally important to AI development and innovation. Contemporary AI and ML systems are trained by ingesting enormous quantities of data. This means their benefits are dependent on the quantity and quality of data. As such, the development and maintenance of datasets is an important priority for Singapore. Government policies are needed to facilitate the access and sharing of data to enable the further development of AI in a collaborative and international manner. One way governments can do this is by developing a unified data policy that allows researchers across agencies access the same high-quality data via a centralized and secured data lake. It would also be good to survey Singapore datasets that are currently available in various key sectors such as transport, healthcare and the economy, and to make an effort to develop representative, unbiased, contextually aware datasets in areas currently lacking.

CONTRIBUTORS

This report is a result of tireless work by contributors from diverse disciplines. The core FRC Study Team whom we credited at the start of this report was supported by over 30 other contributors from across disciplines such as Computer Science, Psychology, Sociology and Law, plus government agency heads. The constructive and valuable feedback of the NRF Scientific Advisory Board (SAB) and the Committee of Government Scientific Advisors (CGSA) has helped make this report more comprehensive. Also making invaluable contributions were administrative contributors such as editors and reviewers. This report would not have been possible without all of them.

Technical

Milad ABDOLLAHZADEH
Research Fellow, Singapore University of Technology and Design

Jennifer ANG
Associate Professor of Philosophy, Singapore University of Social Sciences

Kai Keng ANG
Associate Professor, School of School of Computer Science and Engineering, Nanyang Technological University; Group Leader, Senior Scientist, Institute for Infocomm Research, Agency for Science, Technology and Research

An BO
President's Council Chair Associate Professor of Computer Science, Nanyang Technological University

Gary CHAN
Professor of Law, Singapore Management University

Keshigeyan CHANDRASEGARAN
Researcher, Singapore University of Technology and Design

Simon CHESTERMAN
Dean and Provost's Chair Professor of Law, National University of Singapore

Gao CONG
Professor of Computer Science, Nanyang Technological University

David DE CREMER
Provost's Chair & Professor of Business, National University of Singapore

Mark FINDLAY
Professorial Research Fellow; Director, Centre for AI and Data Governance, Singapore Management University

Gerard GOGGIN
Professor of Communication Studies, Nanyang Technological University

Abhishek GUPTA
Scientist and Technical Lead @SIMTech, Agency for Science, Technology and Research

Yu HAN
Assistant Professor, School of Computer Science and Engineering, Nanyang Technological University

Kotaro HARA
Assistant Professor of Computing, Singapore Management University

Seng Beng HO
Senior Scientist, Institute of High Performance Computing, Agency for Science, Technology and Research

David HSU
Provost's Chair Professor in Computer Science, National University of Singapore

Jing JIANG
Director of AI & Data Science Cluster, Singapore Management University

Kenneth KWOK
Principal Research Scientist, Institute of High Performance Computing, Agency for Science, Technology and Research

Kwan Min LEE
Professor of Communication & Information, Nanyang Technological University

Camilo LIBEDINSKY
Assistant Professor, Department of Psychology, National University of Singapore

Joo Hwee LIM
Principal Scientist II, Agency for Science, Technology and Research

Malika MEGHJANI
Assistant Professor of Information Systems, Singapore University of Technology and Design

Chun Yan MIAO
President’s Chair Professor of Computer Science, Nanyang Technological University

Steven MILLER
Professor Emeritus of Information Systems, Singapore Management University

Desmond ONG
Assistant Professor, School of Computing, National University of Singapore

Zheng SHOU
Assistant Professor of Computer Engineering, National University of Singapore

Rosa SO
Division Head, Institute for Infocomm Research, Agency for Science, Technology and Research

Harold SOH
Assistant Professor of Computer Science, National University of Singapore

Hallam STEVENS
Co-Director of AI Research Institute, Nanyang Technological University

Bernard TAN
Director of Strategy, Planning & Engagement, National Supercomputing Centre, Singapore

U-Xuan TAN
Associate Professor of Engineering, Singapore University of Technology and Design

Ivor TSANG
Director, Centre for Frontier AI Research, Agency for Science, Technology and Research

Iuna TSYRULNEVA
Postdoctoral Fellow, NTU Institute of Science and Technology for Humanity, Nanyang Technological University

Zhaoxia WANG
Senior Scientist, Principal Investigator and Group Manager, Centre for Frontier AI Research, Agency for Science, Technology and Research

Yonggang WEN
President's Chair Professor of Computer Science, Nanyang Technological University

Zee Kin YEONG
Assistant Chief Executive, Infocomm Media Development Authority

Mengmi ZHANG
Research Scientist and Principal Investigator, Institute for Infocomm Research, Agency for Science, Technology and Research

Joey Tianyi ZHOU
Senior Scientist, Centre for Frontier AI Research, Agency for Science, Technology and Research

Administrative

Hui Lin GOH
Head of Strategic Planning & Programmes, Agency for Science, Technology and Research

Vasanth SESHADRI
Editor of the Report; Founder and Creative Director of The Sunny Side Advertising

Ray LIM
Former Research Fellow, Nanyang Technological University

We would also like to thank the following **International External Reviewers**:

Hussein ABBASS
Professor, School of Engineering and Information Technology, UNSW Canberra

Masashi SUGIYAMA
Professor, Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, The University of Tokyo

Dacheng TAO
Professor of Computer Science, School of Computer Science, The University of Sydney

Michael WOOLDRIDGE
Professor of Computer Science, Department of Computer Science, University of Oxford

REFERENCES

1: “AIs 10 to Watch: The Future of AI”. https://ieeexplore.ieee.org/document/8355886

2: “The Future of AI: AI’s 10 to Watch”. https://ieeexplore.ieee.org/document/9309122

3: “RIE2025 Plan”. https://www.nrf.gov.sg/rie2025-plan

4: “Doctor Covid won multiple awards”. https://www.a-star.edu.sg/ihpc/news/news/publicity-highlights/doctor-covid-won-multiple-awards

5: “S'pore's health science innovations get AI boost in SingHealth, SGInnovate tie-up”. (2021). https://www.straitstimes.com/singapore/spores-health-science-innovations-get-boost-from-artificial-intelligence-in-singhealth

6: “AI-powered tool detects lung infection in chest X-rays quickly during Covid-19 screening”. https://www.straitstimes.com/singapore/ai-powered-tool-detects-lung-infection-in-chest-x-rays-quickly-during-covid-19-screening

7: “New app helps pre-diabetics check their food”. https://news.nus.edu.sg/new-app-helps-pre-diabetics-check-their-food/

8: “NTU, TTSH scientists develop glaucoma diagnosis system powered by artificial intelligence”. (2021). https://www.channelnewsasia.com/singapore/glaucoma-diagnosis-ai-system-ttsh-ntu-2160686

9: “Exclusive: Tan Tock Seng Hospital builds “artificial brain” to manage services”. https://govinsider.asia/smart-gov/exclusive-tan-tock-seng-hospital-builds-artificial-brain-to-manage-services/

10: “Woodlands Health Campus to use AI and robotics for patient care”. https://govinsider.asia/innovation/woodlands-health-campus-to-use-ai-and-robotics-for-patient-care/

11: “How will robots help run Singapore’s hospitals?”. https://govinsider.asia/digital-gov/how-will-robots-help-run-singapores-hospitals/

12: “Artificial intelligence-based apps help special needs students learn about emotions”. https://www.straitstimes.com/singapore/artificial-intelligence-based-apps-help-special-needs-students-learn-about-emotions

13: “MAE Professor Chen I-Ming and his team created a disinfection robot to support the fight against COVID-19”. https://www.ntu.edu.sg/mae/news-events/news/detail/mae-professor-chen-i-ming-and-his-team-created-a-disinfection-robot-to-support-the-fight-against-covid-19

14: “Speech Lab”. https://aisingapore.org/speech-lab/

15: Shi, K., Tan, K.M., Duan, R., Salleh, S.U.M., Suhaimi, N.F.A., Vellu, R., Thai, N.T.H.H., Chen, N.F. (2020) Computer-Assisted Language Learning System: Automatic Speech Evaluation for Children Learning Malay and Tamil. Proc. Interspeech 2020, 1019-1020

16: “The 10 Best AI and Data Science Undergraduate Courses for 2021”. https://www.forbes.com/sites/bernardmarr/2020/07/13/the-10-best-ai-and-data-science-undergraduate-courses-for-2021

17: “Andrew Ng predicts the next 10 years in AI”. https://venturebeat.com/ai/andrew-ng-predicts-the-next-10-years-in-ai/

18: “Stanford University Human-Centered Artificial Intelligence”. https://hai.stanford.edu/

19: “Pretty much anything that a normal person can do in <1 sec, we can now automate with AI.” Andrew Ng. https://twitter.com/AndrewYNg/status/788548053745569792

20: J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018. [Online]. Available: https://arxiv.org/abs/1810.04805.

21: T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child and A. Ramesh, “Language Models are Few-Shot Learners,” [Online]. Available: https://arxiv.org/abs/2005.14165.

22: “China’s GPT-3? BAAI Introduces Superscale Intelligence Model ‘Wu Dao 1.0’”. https://syncedreview.com/2021/03/23/chinas-gpt-3-baai-introduces-superscale-intelligence-model-wu-dao-1-0/

23: I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Nets,” Advances in Neural Information Processing Systems, vol. 27, 2014.

24: Dhariwal, P., and Nichol, A., 2021. Diffusion models beat GANs on image synthesis. https://doi.org/10.48550/arXiv.2105.05233

25: “What is Field-weighted Citation Impact (FWCI)?”. https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/~/what-is-field-weighted-citation-impact

26: “Research and Publishing Support: Citation Count”. https://researchguides.smu.edu.sg/researchsupport/citation

27: “Research Metrics: Scholarly Output”. https://libguides.usc.edu.au/c.php?g=508927&p=3480472

28: “National Artificial Intelligence Strategy: Advancing Our Smart Nation Journey”. https://www.smartnation.gov.sg/files/publications/national-ai-strategy.pdf

29: “About AI Singapore”. https://aisingapore.org/about-us/

30: “AI Apprenticeship Programme (AIAP)”. https://aisingapore.org/industryinnovation/aiap/

31: “AI Student Outreach Programme”. https://aisingapore.org/student-outreach-programme/

32: “Centre for Frontier AI Research (CFAR)”. https://www.a-star.edu.sg/cfar/

33: “Advanced Manufacturing and Engineering”. https://www.a-star.edu.sg/i2r/research-capabilities/advance-manufacturing-engineering

34: “CNRS@CREATE”. https://www.cnrsatcreate.cnrs.fr/

35: “DESCARTES: A CNRS@CREATE Program on Intelligent Modelling for Decision-making in Critical Urban Systems”. https://www.cnrsatcreate.cnrs.fr/descartes/

36: “Singtel Cognitive and Artificial Intelligence Lab (SCALE@NTU)”. https://www.ntu.edu.sg/scale

37: “Alibaba-NTU Singapore Joint Research Institute”. https://www.ntu.edu.sg/alibaba-ntu-jri

38: “Institute of Data Science”. https://ids.nus.edu.sg/

39: “SAIL in Sea: A challenge-driven AI research team”. https://sail.sea.com/

40: “A boost for AI research and education with NUSAIL and Rosetta”. https://news.nus.edu.sg/a-boost-for-ai-research-and-education-with-nusail-and-rosetta/

41: “Living Analytics Research Centre”. https://larc.smu.edu.sg/

42: “Collaborative, Robust & Explainable AI-based Decision-making Lab”. https://care-ai.smu.edu.sg/

43: “Centre for AI and Data Governance”. https://caidg.smu.edu.sg/

44: “SMU-A*STAR Joint Lab in Social and Human-Centered Computing”. https://site.smu.edu.sg/sajl

45: “Welcome to the World of Artificial Intelligence and Data Science with SUTD”. https://ai.sutd.edu.sg/

46: “4 Thrusts of AI/DS (SUTD Internal Focus)”. https://ai.sutd.edu.sg/research/

47: “Salesforce AI Research Team Opens First International Hub in Singapore”. https://www.salesforce.com/blog/ai-research-team-singapore/

48: “WHAT THEY ARE SAYING: White House Blueprint for an AI Bill of Rights Lauded as Essential Step Toward Protecting the American Public”. https://www.whitehouse.gov/ostp/news-updates/2022/10/17/what-they-are-sayingwhite-house-blueprint-for-an-ai-bill-of-rights-lauded-as-essential-step-toward-protecting-the-american-public/

49: “NSF partnerships expand National AI Research Institutes to 40 states”. https://beta.nsf.gov/news/nsf-partnerships-expand-national-ai-research-institutes-40-states

50: “New University-wide institute to integrate natural, artificial intelligence”. https://news.harvard.edu/gazette/story/2021/12/new-harvard-institute-to-study-natural-artificial-intelligence/

51: “China’s New AI Governance Initiatives Shouldn’t Be Ignored”. https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127

52: “Beijing Academy of Artificial Intelligence”. https://www.baai.ac.cn/english.html

53: “Wu Dao 2.0 - Bigger, Stronger, Faster AI From China”. https://www.forbes.com/sites/alexzhavoronkov/2021/07/19/wu-dao-20bigger-stronger-faster-ai-from-china/?sh=6aaa174e6fb2

54: “The next frontier for AI in China could add $600 billion to its economy”. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-next-frontier-for-ai-in-china-could-add-600-billion-to-its-economy

55: “AI Strategy 2022”. https://www8.cao.go.jp/cstp/ai/aistratagy2022en_ov.pdf

56: “On Artificial Intelligence – A European approach to excellence and trust”. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

57: “National AI Strategy”. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

58: “The Alan Turing Institute”. https://www.turing.ac.uk/about-us

59: “Emmanuel Macron Talks to WIRED About France's AI Strategy”. https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/

60: “Germany’s human-centred approach to AI is inclusive, evidence-based and capacity-building”. https://oecd.ai/en/wonk/germany-takes-an-inclusive-and-evidence-based-approach-for-capacity-building-and-a-human-centred-use-of-ai

61: “Switzerland – A Trusted Hub for Artificial Intelligence (AI)”. https://www.s-ge.com/sites/default/files/publication/free/factsheet-artificial-intelligence-switzerland-s-ge-en-2022.pdf

62: “Novartis and Microsoft announce collaboration to transform medicine with artificial intelligence”. https://www.novartis.com/news/media-releases/novartis-and-microsoft-announce-collaboration-transform-medicine-artificial-intelligence

63: “Announcing Google Research, Europe”. https://blog.google/around-the-globe/google-europe/announcing-google-research-europe/

64: “The National Artificial Intelligence Centre is launched”. https://www.industry.gov.au/news/the-national-artificial-intelligence-centre-is-launched

65: “Computational Intelligence” by David L. Poole. https://archive.org/details/computationalint00pool

66: S. J. Russell and P. Norvig, Artificial intelligence: A modern approach. USA: Prentice-Hall, Inc., 2010.

67: “What is AI?”. http://www-formal.stanford.edu/jmc/whatisai.pdf

68: “Preparing for the future of Artificial Intelligence”. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

69: “National AI Strategy”. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

70: Nils J. Nilsson, The Quest for Artificial Intelligence: A History of Ideas and Achievements (Cambridge, UK, 2010), Used by Stanford’s report One Hundred Year Study on Artificial Intelligence, 2021

71: A 20-Year Community Roadmap for Artificial Intelligence Research in the US. Computing Community Consortium (CCC) and AAAI, 2019

72: Economist Intelligence Unit.

73: “Google’s Demis Hassabis – misuse of artificial intelligence ‘could do harm’”. https://www.bbc.com/news/business-34266425

74: IBM.

75: Ben Schneiderman. Human-Centered AI, 2022, Oxford University Press, Oxford, UK.

76: “Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk”. https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281

77: “Elon Musk’s regulatory woes mount as U.S. moves closer to recalling Tesla’s self-driving software”. https://fortune.com/2022/06/10/elon-musk-tesla-nhtsa-investigation-traffic-safety-autonomous-fsd-fatal-probe/

78: Ross C, Swetlitz I: IBM’s Watson supercomputer recommended “unsafe and incorrect” cancer treatments, internal documents show. STAT. July 25, 2018. Available at www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

79: Mehrabi, N., Morstatter, F., Saxena, N., et al. (2022) A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54 (6)

80: Kearns, Michael. “Data Intimacy, Machine Learning and Consumer Privacy.” University of Pennsylvania Law School, May 2018 https://www.law.upenn.edu/live/files/7952-kearns-finalpdf

81: Eric Schmidt, “AI, Great Power Competition & National Security, “ Daedulus Journal of the American Academy of Arts and Sciences, Special issue on AI & Society, Volume 151, Number 2; Spring 2022.

82: “The Supply of Disinformation Will Soon Be Infinite”. https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/

83: “How AI Writing Tools Are Helping Students Fake Their Homework”. https://www.lifewire.com/how-ai-writing-tools-are-helping-students-fake-their-homework-6743857

84: “Prime Minister Jacinda Ardern Commencement Address”. https://www.harvardmagazine.com/2022/05/commencement-2022-jacinda-ardern-address

85: “Empowering impactful responsible AI practices”. https://www.microsoft.com/en-us/ai/responsible-ai

86: “Artificial Intelligence at Google: Our Principles”. https://ai.google/principles/

87: “How will Singapore ensure responsible AI use?”. https://govinsider.asia/digital-gov/achim-granzen-forrester-ai-drives-the-evolution-of-technology-and-data-governance/

88: “ANNEX A: Council Members of the Advisory Council on the Ethical use of AI and Data”. https://www.imda.gov.sg/-/media/Imda/Files/About/Media-Releases/2018/Annex-A---Council-Members-of-the-Advisory-Council-on-the-Ethical-use-of-AI-and-Data.pdf

89: “Model Artificial Intelligence Governance Framework”. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

90: “Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework”. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGAIGovUseCases.pdf

91: “Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework, Volume 2”. https://file.go.gov.sg/ai-gov-use-cases-2.pdf

92: “Artificial Intelligence Ethics and Governance Body of Knowledge (AI E&G BoK)”. https://ai-ethics-bok.scs.org.sg/

93: “CET797 Body of Knowledge (BoK) for AI Ethics and Governance”. https://www.ntu.edu.sg/pace/programmes/detail/cet797-body-of-knowledge-(bok)-for-ai-ethics-and-governance

94: Monetary Authority of Singapore (2018) Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector. https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

95: “Singapore’s AI verification framework welcomed by industry”. https://www.itnews.asia/news/singapores-ai-verification-framework-welcomed-by-industry-580658

96: Singapore Management University. External Research Grant Awards. https://research.smu.edu.sg/externalresearchgrants

97: “AI and Compute”. https://openai.com/blog/ai-and-compute/

98: "Sustainable AI: AI for sustainability and the sustainability of AI”. https://d-nb.info/1231344113/34

99: “Tiny Machine Learning at MIT”. https://tinyml.mit.edu/

100: “Edge Intelligence: Edge Computing and Machine Learning (2022 Guide)”. https://viso.ai/edge-ai/edge-intelligence-deep-learning-with-edge-computing

101: “Take Action for the Sustainable Development Goals” https://www.un.org/sustainabledevelopment/sustainable-development-goals/

102: Vinuesa, R., Azizpour, H., Leite, I., et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020).

103: “AI for Sustainability: An overview of AI and the SDGs to contribute to the European policy-making”. https://ec.europa.eu/futurium/en/system/files/ged/vincent-pedemonte_ai-for-sustainability_0.pdf

104: WEF, 2021. Sustainable Development. https://wef-ai.s3.amazonaws.com/WEF_Empowering-AI-Leadership_Sustainable-Development.pdf

105: Rajarethinam, J., Joel, A., Jing, T. (2020). The influence of South East Asia Forest Fires on ambient Particulate Matter concentrations in Singapore: An Ecological Study using Random Forest and Vector Autoregressive Models. 10.21203/rs.3.rs-41720/v1.

106: Rolnick, D., Donti, P., Kaack, L. (2019) Tackling Climate Change with Machine Learning. arXiv:1906.05433

107: Larsson, S., Anneroth, M., Fellander, A., et al. (2019) Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence. AI Sustainability Center

108: Kindylidi, I., Cabral, T.S. (2021) Sustainability of AI: The case of provision of information to consumers. Sustainability, 13 (21), p. 12064.

109: Ong, Y. S., Gupta, A. (2019). AIR 5: Five pillars of artificial intelligence research. IEEE Transactions on Emerging Topics in Computational Intelligence, 3(5), 411-415.

110: “AlphaFold: a solution to a 50-year-old grand challenge in biology”. https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

111: V. Belle and I. Papantonis, Principles and practice of explainable machine learning, Front. Big Data, vol. 4, no. July, pp. 1–25, 2021, doi: 10.3389/fdata.2021.688969.

112: R. J. Wallace, Three conceptions of rational agency, Ethical Theory Moral Pract., vol. 2, no. 3, pp. 217–242, 1999.

113: S. J. Russell, Rationality and intelligence, Artif. Intell., vol. 94, no. 1–2, pp. 57–77, 1997, doi: 10.1017/9781108770422.047.

114: S.-B. Ho and Z. Wang, Language and robotics: Complex sentence understanding, in International Conference on Intelligent Robotics and Applications, 2019, vol. 11745 LNAI, no. May 2020, pp. 641–654, doi: 10.1007/978-3-030-27529-7_54.

115: S.-B. Ho and Z. Wang, On true language Understanding, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11633 LNCS, no. August, pp. 87–99, doi: 10.1007/978-3-030-24265-7_8.

116: R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in Proc. 21st

ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2015, pp. 1721–1730.

117: A. Karpatne et al., “Theory-guided data science: A new paradigm for scientific discovery from data,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2318–2331, Oct. 2017.

118: X. Jia et al., “Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles,” in Proc. SIAM Int. Conf. Data Mining, May 2019, pp. 558–566.

119: Susskind, Z., Arden, B., John, L.K., Stockton, P. and John, E.B., 2021. Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization. https://doi.org/10.48550/arXiv.2109.06133

120: Graziani, M., Dutkiewicz, L., Calvaresi, D. et al. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Artif Intell Rev (2022). https://doi.org/10.1007/s10462-022-10256-8

121: Prashant Gohel, Priyanka Singh, Manoranjan Mohanty. Explainable AI: current status and future directions. (2021).

122: “Explainable AI as a Service for Community Healthcare”. https://aisingapore.org/explainable-ai-as-a-service/

123: Varshneya, Saurabh; Ledent, Antoine; Vandermuelen, Rob; Lei, Yunwen; Enders, Matthias; Borth, Damian; and Kloft, Marius. Learning interpretable concept groups in CNNs. (2021). Proceedings of the 30th International Joint Conference on Artificial Intelligence, Montreal, 2021 August 19-27. 1061-1067. Research Collection School Of Computing and Information Systems. Available at: https://ink.library.smu.edu.sg/sis_research/7206

124: E. Tjoa and C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Networks Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

125: “Stable Diffusion Public Release”. https://stability.ai/blog/stable-diffusion-public-release

126: Chenfei Wu, et al. "Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis." arXiv preprint arXiv:2207.09814 (2022).

127: Ruben Villegas, et al. "Phenaki: Variable length video generation from open domain textual description." arXiv preprint arXiv:2210.02399 (2022).

128: Ben Poole, et al. "Dreamfusion: Text-to-3d using 2d diffusion." arXiv preprint arXiv:2209.14988 (2022).

129: Rongjie Huang, et al. "Prodiff: Progressive fast diffusion model for high-quality text-to-speech." Proceedings of the 30th ACM International Conference on Multimedia. 2022.

130: Uriel Singer, et al. "Make-a-video: Text-to-video generation without text-video data." arXiv preprint arXiv:2209.14792 (2022).

131: “DeepMind’s AlphaCode AI writes code at a competitive level”. https://techcrunch.com/2022/02/02/deepminds-alphacode-ai-writes-code-at-a-competitive-level

132: “5 AI Tools that can Generate Code to Help Programmers”. https://www.forbes.com/sites/janakirammsv/2022/03/14/5-ai-tools-that-can-generate-code-to-help-programmers/

133: Yamins DL et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences. 2014 Jun 10;111(23):8619-24.

134: Dabney W et al. A distributional code for value in dopamine-based reinforcement learning. Nature. 2020 Jan;577(7792):671-5.

135: Schrimpf M et al. The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences. 2021 Nov 9;118(45).

136: Dorkenwald S et al. Automated synaptic connectivity inference for volume electron microscopy. Nature methods. 2017 Apr;14(4):435-42.

137: Mathis A et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience. 2018 Sep;21(9):1281-9.

138: Frey M, Nau M, Doeller CF. Magnetic resonance-based eye tracking using deep neural networks. Nature neuroscience. 2021 Dec;24(12):1772-9.

139: Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. Nature. 2019 Apr;568(7753):493-8.

140: Willett FR et al. High-performance brain-to-text communication via handwriting. Nature. 2021 May;593(7858):249-54.

141: Peratham Wiriyathammabhum, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloimonos. 2016. Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics. ACM Comput. Surv. 49, 4, Article 71 (December 2017), 44 pages. https://doi.org/10.1145/3009906

142: Duan, Jiafei and Yu, Samson and Tan, Hui Li and Zhu, Hongyuan and Tan, Cheston. A Survey of Embodied AI: From Simulators to Research Tasks. 2021 Mar 8. https://arxiv.org/abs/2103.04918

143: Andrew Jaegle, et al. "Perceiver: General perception with iterative attention." International conference on machine learning. PMLR, 2021.

144: Jean-Baptiste Alayrac, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

145: Michael Ahn, et al. "Do as i can, not as i say: Grounding language in robotic affordances." arXiv preprint arXiv:2204.01691 (2022).

146: “Towards Helpful Robots: Grounding Language in Robotic Affordances”. https://ai.googleblog.com/2022/08/towards-helpful-robots-grounding.html

147: B. Green and Y. Chen, “Disparate interactions: An algorithm-in-the- loop analysis of fairness in risk assessments,” in Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 90– 99.

148: I. D. Raji and J. Buolamwini, “Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products,” in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

149: T.Schnabel, A.Swaminathan, A.Singh, N.Chandak, and T.Joachims, “Recommendations as treatments: Debiasing learning and evaluation,” in international conference on machine learning. PMLR, 2016, pp. 1670–1679.

150: J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, “Bias and debias in recommender system: A survey and future directions,” arXiv preprint arXiv:2010.03240, 2020.

151: Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

152: S. Caton and C. Haas, “Fairness in machine learning: A survey,” arXiv preprint arXiv:2010.04053, 2020.

153: R.K.Bellamy, K.Dey, M.Hind ,S.C.Hoffman, S.Houde, K.Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic et al., “Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” arXiv preprint arXiv:1810.01943, 2018.

154: F. Kamiran and T. Calders, “Data preprocessing techniques for classi- fication without discrimination,” Knowledge and information systems, vol. 33, no. 1, pp. 1–33, 2012.

155: M.Feldman,S.A.Friedler,J.Moeller,C.Scheidegger,andS.Venkatasubramanian, “Certifying and removing disparate impact,” in proceed- ings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

156: H. Jiang and O. Nachum, “Identifying and correcting label bias in ma- chine learning,” in International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 702–712.

157: M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” in Artificial Intelligence and Statistics. PMLR, 2017, pp. 962–970.

158: J. Komiyama, A. Takeda, J. Honda, and H. Shimao, “Nonconvex optimization for regression with fairness constraints,” in International conference on machine learning. PMLR, 2018, pp. 2737–2746.

159: I. Valera, A. Singla, and M. Gomez Rodriguez, “Enhancing the accuracy and fairness of human decision making,” Advances in Neural Information Processing Systems, vol. 31, 2018.

160: Alvin Chan, Nanyang Technological University (NTU), Defences and Threats in Safe Deep Learning. (2021).

161: A.Madry, A.Makelov, L.Schmidt, D.Tsipras, and A.Vladu,“Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.

162: H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” arXiv preprint arXiv:1901.08573, 2019.

163: A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, “Adversarial training for free!” arXiv preprint arXiv:1904.12843, 2019.

164: J.Zhang, X.Xu, B.Han, G.Niu, L.Cui, M.Sugiyama and M.Kankanhalli, “Attacks which do not kill training make adversarial learning stronger,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 278–11 287.

165: H. Drucker and Y. Le Cun, “Double backpropagation increasing generalization performance,” in IJCNN-91-Seattle International Joint Conference on Neural Networks, vol. 2. IEEE, 1991, pp. 145–150.

166: A. S. Ross and F. Doshi-Velez, “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients,” in Thirty-second AAAI conference on artificial intelligence, 2018.

167: D. Jakubovitz and R. Giryes, “Improving dnn robustness to adversarial attacks using jacobian regularization,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 514–529.

168: C. Etmann, S. Lunz, P. Maass, and C.-B. Scho ̈nlieb, “On the connection between adversarial robustness and saliency map interpretability,” arXiv preprint arXiv:1905.04172, 2019.

169: A. Chan, Y. Tay, Y. S. Ong, and J. Fu, “Jacobian adversarially regularized networks for robustness,” arXiv preprint arXiv:1912.10185, 2019.

170: A. Chan, Y. Tay, and Y.-S. Ong, “What it thinks is important is important: Robustness transfers through input gradients,” in Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 332–341.

171: M.Hein and M.Andriushchenko, “Formal guarantees on the robustness of a classifier against adversarial manipulation,” in Advances in Neural Information Processing Systems, 2017, pp. 2266–2276.

172: A. Raghunathan, J. Steinhardt, and P. S. Liang, “Semidefinite relaxations for certifying robustness to adversarial examples,” in Advances in Neural Information Processing Systems, 2018, pp. 10 877–10 887.

173: J.Cohen, E.Rosenfeld and Z.Kolter, “Certified adversarial robustness via randomized smoothing,” in International Conference on Machine Learning. PMLR, 2019, pp. 1310–1320.

174: M. Balunovic and M. Vechev, “Adversarial training and provable defenses: Bridging the gap,” in International Conference on Learning Representations, 2019.

175: A. Blum, T. Dick, N. Manoj, and H. Zhang, “Random smoothing might be unable to certify l∞ robustness for high-dimensional images,” Journal of Machine Learning Research, vol. 21, pp. 1–21, 2020.

176: B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar, and K. Xia, “Exploiting machine learning to subvert your spam filter.” LEET, vol. 8, pp. 1–9, 2008.

177: H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, “Support vector machines under adversarial label contamination,” Neu- rocomputing, vol. 160, pp. 53–62, 2015.

178: J. Steinhardt, P. W. W. Koh, and P. S. Liang, “Certified defenses for data poisoning attacks,” in Advances in neural information processing systems, 2017, pp. 3517–3529.

179: T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnera- bilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017.

180: Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, “Trojaning attack on neural networks,” in 25nd Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018. The Internet Society, 2018.

181: Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, “Turning your weakness into a strength: Watermarking deep neural networks by backdooring,” in 27th {USENIX} Security Symposium ({USENIX} Security 18), 2018, pp. 1615–1631.

182: A. Chan, Y. Tay, Y.-S. Ong, and A. Zhang, “Poison attacks against text datasets with conditional adversarially regularized autoencoder,” arXiv preprint arXiv:2010.02684, 2020.

183: B.Tran, J.Li, and A.Madry, “Spectral signatures in backdoor attacks,” in Advances in Neural Information Processing Systems, 2018, pp. 8011–8021.

184: K. Liu, B. Dolan-Gavitt, and S. Garg, “Fine-pruning: Defending against backdooring attacks on deep neural networks,” in International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, 2018, pp. 273–294.

185: Y. Ma, X. Zhu, and J. Hsu, “Data poisoning against differentially- private learners: Attacks and defenses,” arXiv preprint arXiv:1903.09860, 2019.

186: E. Rosenfeld, E. Winston, P. Ravikumar, and Z. Kolter, “Certified robustness to label-flipping attacks via randomized smoothing,” in International Conference on Machine Learning. PMLR, 2020, pp. 8230–8241.

187: M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li, “Rab: Provable robustness against backdoor attacks,” arXiv preprint arXiv:2003.08904, 2020.

188: S. Dodge and L. Karam, “A study and comparison of human and deep learning recognition performance under visual distortions,” in 2017 26th international conference on computer communication and networks (ICCCN). IEEE, 2017, pp. 1–7.

189: D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt and D. Song,“Natural adversarial examples,” arXiv preprint arXiv:1907.07174, 2019.

190: E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation policies from data,” arXiv preprint arXiv:1805.09501, 2018.

191: D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” arXiv preprint arXiv:2006.16241, 2020.

192: K. Kireev, M. Andriushchenko, and N. Flammarion, “On the effec- tiveness of adversarial training against common corruptions,” arXiv preprint arXiv:2103.02325, 2021.

193: Stanford University, 2021. AI Index Report. https://aiindex.stanford.edu/report/

194: The State of AI Ethics, Montreal AI Ethics Institute, 2021 https://montrealethics.ai/wp-content/uploads/2021/01/The-State-of-AI-Ethics-Report-January-2021.pdf

195: FICO, 2021. The State of Responsible AI: 2021. https://www.fico.com/en/latest-thinking/market-research/state-responsible-ai-2021

196: The Economist Intelligence Unit. “Staying ahead of the curve: The business case for responsible AI”. https://www.eiu.com/n/staying-ahead-of-the-curve-the-business-case-for-responsible-ai/

197: De Cremer, D., McGuire, J. (2021) Human–Algorithm Collaboration Works Best if Humans Lead (Because it is Fair!). Soc Just Res. https://doi.org/10.1007/s11211-021-00382-z

198: Miller, S. (2018) AI: Augmentation, more so than automation. Asian Management Insights. 5, (1), 1-20.

199: Davenport, T., Miller, S. (2020) The future of work now: AI-driven transaction surveillance at DBS Bank. https://www.forbes.com/sites/tomdavenport/2020/10/23/the-future-of-work-now-ai-driven-transaction-surveillance-at-dbs-bank

200: Lee, S., Miller, S. (2019) AI gets real at Singapore's Changi Airport (Part 1). Asian Management Insights. 6, (1), 10-19.

201: Davenport, T., Miller, S. (2020) The future of work now: The multi-faceted mall security guard at a multi-faceted Jewel. https://www.forbes.com/sites/tomdavenport/2020/09/28/the-future-of-work-now-the-multi-faceted-mall-security-guard-at-a-multi-faceted-jewel

202: Stevens, H., Vale, D. (2021) A New AI Lexicon: Smart. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-smart-b6be6d38bace

203: “Federated Learning: Opportunities and Challenges”: https://arxiv.org/abs/2101.05428

204: “A Survey on Federated Learning”: https://www.researchgate.net/publication/344871928_A_Survey_on_Federated_Learning_The_Journey_From_Centralized_to_Distributed_On-Site_Learning_and_Beyond

205: J. M. Drazen, S. Morrissey, D. Malina, M. B. Hamel, and E. W. Campion. The importance — and the complexities — of data sharing. New England Journal of Medicine, 375(12):1182–1183, 2016.

206: H. M. Krumholz and J. Waldstreicher. Toward fairness in data sharing. New England Journal of Medicine, 375(5):405–407, 2016.

207: Ocean Protocol Foundation. Ocean protocol: A decentralized substrate for AI data and services. Technical whitepaper, 2019.

208: N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso. The future of digital health with federated learning. npj Digital Medicine, 3(119), 2020.

209: D. Ng, X. Lan, M. M.-S. Yao, W. P. Chan, and M. Feng. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quantitative Imaging in Medicine and Surgery, 11(2), 2020.

210: D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau. Federated learning for keyword spotting. In Proc. ICASSP, 2019.

211: A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. arXiv:1811.03604, 2018.

212: Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics. PMLR, 2938–2948.

213: Q. M. Hoang, T. N. Hoang, B. K. H. Low, and C. Kingsford. Collective model fusion for multiple black-box experts. In *Proc. ICML*, pages 2742-2750, 2019.

214: T. N. Hoang, C. T. Lam, B. K. H. Low, and P. Jaillet. Learning task-agnostic embedding of multiple black-box experts for multi-task model fusion. In *Proc. ICML*, pages 4282-4292, 2020.

215: C. T. Lam, T. N. Hoang, B. K. H. Low, and P. Jaillet. Model Fusion for Personalized Learning. In *Proc. ICML*, pages 5948-5958, 2021.

216: A. Ghorbani and J. Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proc. ICML*, pages 2242–2251, 2019.

217: A. Ghorbani, M. P. Kim, and J. Zou. A distributional framework for data valuation. In *Proc. ICML*, 2020.

218: R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. In *Proc. VLDB Endowment*, pages 1610–1623, 2019a.

219: R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gurel, B. Li, C. Zhang, D. Song, and C. Spanos. Towards efficient data valuation based on the Shapley value. In *Proc. AISTATS*, pages 1167–1176, 2019b.

220: X. Xu, Z. Wu, C. S. Foo, and B. K. H. Low. Validation free and replication robust volume-based data valuation. In *Proc. NeurIPS*, 2021a.

221: J. Yoon, S. O. Arik, and T. Pfister. Data valuation using reinforcement learning. In *Proc. ICML*, 2020.

222: R. H. L. Sim, Y. Zhang, M. C. Chan, and B. K. H. Low. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pages 8927–8936, 2020.

223: S. Tay, X. Xu, C. S. Foo, and B. K. H. Low. Incentivizing collaboration in machine learning via synthetic data rewards. In Proc. AAAI, 2022.

224: X. Xu, L. Lyu, X. Ma, C. Miao, C. S. Foo, and B. K. H. Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. NeurIPS*, 2021b.

225: X. Fan, Y. Ma, Z. Dai, W. Jing, C. Tan, and B. K. H. Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Proc. NeurIPS*, 2021.

226: Z. Dai, B. K. H. Low, and P. Jaillet. Federated Bayesian optimization via Thompson sampling. In *Proc. NeurIPS*, pages 9687–9699, 2020.

227: Z. Dai, B. K. H. Low, and P. Jaillet. Differentially Private Federated Bayesian Optimization with Distributed Exploration. In *Proc. NeurIPS*, 2021.

228: R. H. L. Sim, Y. Zhang, B. K. H. Low, and P. Jaillet. Collaborative Bayesian optimization with fair regret. In *Proc. ICML*, pages 9691–9701, 2021.

229: Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *Proc. IEEE Symposium on Security and Privacy*, pages 463–480, 2015.

230: Y. Chen, S. Zhang, and B. K. H. Low. Near-Optimal Task Selection for Meta-Learning with Mutual Information and Online Variational Bayesian Unlearning. In *Proc. AISTATS*, 2022.

231: Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong Anomaly Detection Through Unlearning. In *Proc. ACM CCS*, pages 1283–1297, 2019.

232: Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI forget you: Data deletion in machine learning. In *Proc. NeurIPS*, pages 3513–3526, 2019.

233: Q. P. Nguyen, B. K. H. Low, and P. Jaillet. Variational Bayesian unlearning. In *Proc. NeurIPS*, pages 16025–16036, 2020.

234: Q. P. Nguyen, R. Oikawa, D. M. Divakaran, M. C. Chan, B. K. H. Low. Markov Chain Monte Carlo-Based Machine Unlearning: Unlearning What Needs to be Forgotten. In Proc. AsiaCCS, 2022.

235: “Trustworthy Federated Ubiquitous Learning (TrustFUL) Research Lab”. https://trustful.federated-learning.org/

236: “Toward Trustable Model-centric Sharing for Collaborative Machine Learning”. https://ids.nus.edu.sg/TrustedCollabML.html

237: B. Goertzel, “Artificial general intelligence: Concept, state of the art, and future prospects,” Journal of Artificial General Intelligence, vol. 5, no. 1, pp. 1–46, 2014.

238: S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

239: K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” J Big Data, 2016.

240: F. Zhuang et al., “A comprehensive survey on transfer learning,” Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

241: Y. Zhang and Q. Yang, “A survey on multi-task learning,” IEEE Transactions on Knowledge and Data Engineering, 2021.

242: M. Crawshaw, “Multi-task learning with deep neural networks: A survey,” CoRR, vol. abs/2009.09796, 2020, Available: https://arxiv.org/abs/2009.09796

243: J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” Proceedings of the national academy of sciences, vol. 114, no. 13, pp. 3521–3526, 2017.

244: C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep network,” in ICML, 2017.

245: S. Thrun and L. Pratt, Learning to learn. Springer Science & Business Media, 2012.

246: T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, “Meta-learning in neural networks: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

247: S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” IEEE transactions on pattern analysis and machine intelligence, 2021.

248: M. Crawshaw, “Multi-task learning with deep neural networks: A survey,” CoRR, vol. abs/2009.09796, 2020, Available: https://arxiv.org/abs/2009.09796

249: S. Liu, Y. Liang, and A. Gitter, “Loss-balanced task weighting to reduce negative transfer in multi-task learning,” in Proceedings of the AAAI conference on artificial intelligence, 2019, vol. 33, pp. 9977–9978.

250: M. Abdollahzadeh, T. Malekzadeh, and N.-M. Cheung, “Revisit multi-modal meta-learning through the lens of multi-task learning,” in International conference on neural information processing systems (NeurIPS-2021), 2021.

251: S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, “Icarl: Incremental classifier and representation learning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2001–2010.

252: D. Rolnick, A. Ahuja, J. Schwarz, T. P. Lillicrap, and G. Wayne, “Experience replay for continual learning,” in Proceedings of the 33rd international conference on neural information processing systems, 2019, pp. 350–360.

253: D. Isele and A. Cosgun, “Selective experience replay for lifelong learning,” in Proceedings of the AAAI conference on artificial intelligence, 2018, vol. 32.

254: M. De Lange and T. Tuytelaars, “Continual prototype evolution: Learning online from non-stationary data streams,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8250–8259.

255: D. Lopez-Paz and M. Ranzato, “Gradient episodic memory for continual learning,” Advances in neural information processing systems, vol. 30, pp. 6467–6476, 2017.

256: A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with a-GEM,” in International conference on learning representations, 2018.

257: H. Shin, J. K. Lee, J. Kim, and J. Kim, “Continual learning with deep generative replay,” in Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 2994–3003.

258: F. Lavda, J. Ramapuram, M. Gregorova, and A. Kalousis, “Continual classification learning using generative model,” in Proceedings of the 32nd conference on neural information processing systems (NeurIPS) 2018, 2018.

259: J. Ramapuram, M. Gregorova, and A. Kalousis, “Lifelong generative modeling,” Neurocomputing, vol. 404, pp. 381–400, 2020.

260: Z. Li and D. Hoiem, “Learning without forgetting,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pp. 2935–2947, 2017.

261: H. Jung, J. Ju, M. Jung, and J. Kim, “Less-forgetting learning in deep neural networks,” arXiv preprint arXiv:1607.00122, 2016.

262: J. Zhang et al., “Class-incremental learning via deep model consolidation,” in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1131–1140.

263: S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, “Overcoming catastrophic forgetting by incremental moment matching,” in Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4655–4665.

264: F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” in International conference on machine learning, 2017, pp. 3987–3995.

265: A. A. Rusu et al., “Progressive neural networks,” arXiv preprint arXiv:1606.04671, 2016.

266: R. Aljundi, P. Chakravarty, and T. Tuytelaars, “Expert gate: Lifelong learning with a network of experts,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3366–3375.

267: J. Xu and Z. Zhu, “Reinforced continual learning,” in Proceedings of the 32nd international conference on neural information processing systems, 2018, pp. 907–916.

268: A. Rosenfeld and J. K. Tsotsos, “Incremental learning through deep adaptation,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 3, pp. 651–663, 2018.

269: C. Fernando et al., “Pathnet: Evolution channels gradient descent in super neural networks,” arXiv preprint arXiv:1701.08734, 2017.

270: A. Mallya and S. Lazebnik, “Packnet: Adding multiple tasks to a single network by iterative pruning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7765–7773.

271: A. Mallya, D. Davis, and S. Lazebnik, “Piggyback: Adapting a single network to multiple tasks by learning to mask weights,” in Proceedings of the european conference on computer vision (ECCV), 2018, pp. 67–82.

272: J. Serra, D. Suris, M. Miron, and A. Karatzoglou, “Overcoming catastrophic forgetting with hard attention to the task,” in International conference on machine learning, 2018, pp. 4548–4557.

273: C. S. Lee and A. Y. Lee, “Clinical applications of continual learning machine learning,” The Lancet Digital Health, vol. 2, no. 6, pp. e279–e281, 2020.

274: W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in Proceedings of the 24th international conference on machine learning, 2007, pp. 193–200. doi: 10.1145/1273496.1273521.

275: Y. Yao and G. Doretto, “Boosting for transfer learning with multiple sources,” in 2010 IEEE computer society conference on computer vision and pattern recognition, 2010, pp. 1855–1862. doi: 10.1109/CVPR.2010.5539857.

276: D. Pardoe and P. Stone, “Boosting for regression transfer,” in Proceedings of the 27th international conference on international conference on machine learning, 2010, pp. 863–870.

277: C. Q. Wan, R. Pan, and J. Li, “Bi-weighting domain adaptation for cross-language text classification,” in IJCAI, 2011.

278: N. Li, H. Hao, Q. Gu, D. Wang, and X. Hu, “A transfer learning method for automatic identification of sandstone microscopic images,” Comput. Geosci., vol. 103, no. C, pp. 111–121, Jun. 2017, doi: 10.1016/j.cageo.2017.03.007.

279: Y. Xu et al., “A unified framework for metric transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 6, pp. 1158–1171, 2017, doi: 10.1109/TKDE.2017.2669193.

280: X. Liu, Z. Liu, G. Wang, Z. Cai, and H. Zhang, “Ensemble transfer learning algorithm,” IEEE Access, vol. 6, pp. 2389–2396, 2018, doi: 10.1109/ACCESS.2017.2782884.

281: E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, “Deep domain confusion: Maximizing for domain invariance,” arXiv preprint arXiv:1412.3474, 2014.

282: M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” 2015, pp. 97–105.

283: A. Gretton et al., “Optimal kernel choice for large-scale two-sample tests,” in Advances in neural information processing systems, 2012, vol. 25. Available: https://proceedings.neurips.cc/paper/2012/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf

284: M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in International conference on machine learning, 2017, pp. 2208–2217.

285: H. Phan et al., “Towards more accurate automatic sleep staging via deep transfer learning,” IEEE Transactions on Biomedical Engineering, vol. 68, no. 6, pp. 1787–1798, 2020.

286: P. Perera and V. M. Patel, “Deep transfer learning for multiple class novelty detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11544–11552.

287: K. Han, A. Vedaldi, and A. Zisserman, “Learning to discover novel visual categories via deep transfer clustering,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8401–8409.

288: N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, “Automated deep transfer learning-based approach for detection of COVID-19 infection in chest x-rays,” IRBM, 2020, doi: https://doi.org/10.1016/j.irbm.2020.07.001.

289: S. Kornblith, J. Shlens, and Q. V. Le, “Do better imagenet models transfer better?” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2661–2671.

290: D. Hendrycks, K. Lee, and M. Mazeika, “Using pre-training can improve model robustness and uncertainty,” in International conference on machine learning, 2019, pp. 2712–2721.

291: Y. Zhong and A. Maki, “Regularizing CNN transfer learning with randomised regression,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13637–13646.

292: Z. Zheng and Y. Yang, “Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation,” International Journal of Computer Vision, vol. 129, no. 4, pp. 1106–1120, 2021.

293: K. Choudhary et al., “Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning,” 2019, doi: https://doi.org/10.1038/s41467-019-13297-w.

294: M. Talo, U. B. Baloglu, Özal Yildirim, and U. R. Acharya, “Application of deep transfer learning for automated brain abnormality classification using MR images,” Cognitive Systems Research, vol. 54, pp. 176–188, 2019.

295: W. Zhenghong, H. Jiang, Z. Ke, and L. Xingqiu, “An adaptive deep transfer learning method for bearing fault diagnosis,” Measurement, vol. 151, p. 107227, Nov. 2019, doi: 10.1016/j.measurement.2019.107227.

296: M. Geng, Y. Wang, T. Xiang, and Y. Tian, “Deep transfer learning for person re-identification,” arXiv preprint arXiv:1611.05244, 2016.

297: D. George, H. Shen, and E. Huerta, “Deep transfer learning: A new deep learning glitch classification method for advanced LIGO,” arXiv preprint arXiv:1706.07446, 2017.

298: M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Unsupervised domain adaptation with residual transfer networks,” in NIPS, 2016, pp. 136–144.

299: H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, “Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1182–1194, 2018, doi: 10.1109/TPAMI.2017.2656884.

300: J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7304–7308, 2013.

301: M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in 2014 IEEE conference on computer vision and pattern recognition, 2014, pp. 1717–1724. doi: 10.1109/CVPR.2014.222.

302: H. Zhu, M. Long, J. Wang, and Y. Cao, “Deep hashing network for efficient similarity retrieval,” 2016, pp. 2415–2421.

303: J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” arXiv preprint arXiv:1411.1792, 2014.

304: L. Mihalkova, T. Huynh, and R. J. Mooney, “Mapping and revising markov logic networks for transfer learning,” in Proceedings of the 22nd national conference on artificial intelligence - volume 1, 2007, pp. 608–614.

305: L. Mihalkova and R. J. Mooney, “Transfer learning from minimal target data by mapping across relational domains,” in Proceedings of the 21st international joint conference on artificial intelligence, 2009, pp. 1163–1168.

306: J. Davis and P. Domingos, “Deep transfer via second-order markov logic,” in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 217–224. doi: 10.1145/1553374.1553402.

307: J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

308: K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

309: X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” arXiv preprint arXiv:2003.04297, 2020.

310: X. Chen and K. He, “Exploring simple siamese representation learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15750–15758.

311: X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, “Dense contrastive learning for self-supervised visual pre-training,” 2021.

312: T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

313: J.-B. Grill, F. Strub, F. Altch´e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar et al., “Bootstrap your own latent: A new approach to self-supervised learning,” arXiv preprint arXiv:2006.07733, 2020.

314: Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10743–10752, 2021.

315: Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In Proceedings of the European Conference on Computer Vision (ECCV), pages 218–234, 2018.

316: Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In CVPR AI for Content Creation Workshop, 2020.

317: Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F.

Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 15885–15896. 2020.

318: Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In CVPR, 2022.

319: Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, Ngai-Man Cheung. Few-shot Image Generation via Adaptation-Aware Kernel Modulation. In NeurIPS-2022.

320: J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in NeurIPS, 2017, pp. 4077–4087.

321: O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., “Matching networks for one shot learning,” in NeurIPS, 2016, pp. 3630–3638.

322: F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1199–1208.

323: W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, “A closer look at few-shot classification,” in ICLR, 2019.

324: S. M. Kye, H. B. Lee, H. Kim, and S. J. Hwang, “Meta-learned confidence for few-shot learning,” arXiv e-prints, pp. arXiv–2002, 2020.

325: A. Ravichandran, R. Bhotika, and S. Soatto, “Few-shot learning with embedded class models and shot-free meta training,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 331–339.

326: A. A. Rusu et al., “Meta-learning with latent embedding optimization,” in International conference on learning representations, 2018.

327: Z. Li, F. Zhou, F. Chen, and H. Li, “Meta-sgd: Learning to learn quickly for few-shot learning,” arXiv preprint arXiv:1707.09835, 2017.

328: S. Baik, S. Hong, and K. M. Lee, “Learning to forget for meta-learning,” in IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2020.

329: M. A. Jamal and G.-J. Qi, “Task agnostic meta-learning for few-shot learning,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11719–11727.

340: Z. Xu, X. Chen, W. Tang, J. Lai, and L. Cao, “Meta weight learning via model-agnostic meta-learning,” Neurocomputing, vol. 432, pp. 124–132, 2021.

241: A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” 2017.

242: B. Hariharan and R. Girshick, “Low-shot visual recognition by shrinking and hallucinating features,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3018–3027.

243: Q. Luo, L. Wang, J. Lv, S. Xiang, and C. Pan, “Few-shot learning via feature hallucination with variational inference,” in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 3963–3972.

244: R. Ni, M. Goldblum, A. Sharaf, K. Kong, and T. Goldstein, “Data augmentation for meta-learning,” in International conference on machine learning, 2021, pp. 8152–8161.

255: S. Qiao, C. Liu, W. Shen, and A. L. Yuille, “Few-shot image recognition by predicting parameters from activations,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7229–7238.

256: H. Qi, M. Brown, and D. G. Lowe, “Low-shot learning with imprinted weights,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5822–5830.

257: S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in CVPR, 2018, pp. 4367–4375.

258: Y. Guo and N.-M. Cheung, “Attentive weights generation for few shot learning via information maximization,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2020, pp. 13499–13508.

259: R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in ACM conference on computer and communications security (CCS), 2015.

260: X. He and Y. Zhang, “Quantifying and mitigating privacy risks of contrastive learning,” in Proc. Of the 2021 ACM SIGSAC conference on computer and communications security, 2021.

261: M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” ACM CCS, 2015.

262: N. Carlini et al., “Extracting training data from large language models,” in 30th USENIX security symposium (USENIX security 21), Aug. 2021, pp. 2633–2650. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

263: R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *IEEE symposium on security and privacy*, 2016.

264: A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, “ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models,” in *Network and distributed system security symposium*, 2019.

265: Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, “The secret revealer: Generative model-inversion attacks against deep neural networks,” *CVPR*, 2020.

266: L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in *IEEE symposium on security and privacy*, 2019, pp. 497–512.

267: C. Song and V. Shmatikov, “Overlearning reveals sensitive attributes,” in International conference on learning representations (ICLR), 2020.

268: A. J. Larrazabala, N. Nieto, V. Peterson, D. H. Milonea, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” Proceedings of the National Academy of Sciences, 2020.

269: Nicolas Kourtellis, Kleomenis Katevas, and Diego Perino. 2020. FLaaS: Federated Learning as a Service. arXiv preprint arXiv:2011.09359 (2020).

270: Swersky, K., Snoek, J., & Adams, R. P. (2013). Multi-task bayesian optimization. Advances in neural information processing systems, 26.

271: Ong, Y. S., & Gupta, A. (2016). Evolutionary multitasking: a computer science view of cognitive multitasking. Cognitive Computation, 8(2), 125-142.

272: Gupta, A., Ong, Y. S., & Feng, L. (2018). Insights on transfer optimization: Because experience is the best teacher. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 51-64.

273: Miikkulainen, R., & Forrest, S. (2021). A biological perspective on evolutionary computation. Nature Machine Intelligence, 3(1), 9-15.

274: Mei, Y., Ardeh, M. A., & Zhang, M. (2021). Knowledge transfer in genetic programming hyper-heuristics. In Automated Design of Machine Learning and Search Algorithms (pp. 149-169). Springer, Cham.

275: Da, B., Gupta, A., & Ong, Y. S. (2019). Curbing negative influences online for seamless transfer evolutionary optimization. IEEE transactions on cybernetics, 49(12), 4365-4378.

276: Ruan, G., Minku, L. L., Menzel, S., Sendhoff, B., & Yao, X. (2020, July). Computational Study on Effectiveness of Knowledge Transfer in Dynamic Multi-objective Optimization. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

277: Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), 148-175.

278: Dai, Z., CHEN, Y., Yu, H., Low, B.K.H. and Jaillet, P., 2022, February. On Provably Robust Meta-Bayesian Optimization. In The 38th Conference on Uncertainty in Artificial Intelligence.

279: Souza, A., Nardi, L., Oliveira, L. B., Olukotun, K., Lindauer, M., & Hutter, F. (2021, September). Bayesian Optimization with a Prior for the Optimum. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 265-296). Springer, Cham.

280: Yogatama, D., & Mann, G. (2014). Efficient transfer learning method for automatic hyperparameter tuning. In Artificial intelligence and statistics (pp. 1077-1085). PMLR.

281: Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013, May). Collaborative hyperparameter tuning. In International conference on machine learning (pp. 199-207). PMLR.

282: Min, A. T. W., Ong, Y. S., Gupta, A., & Goh, C. K. (2019). Multiproblem surrogates: Transfer evolutionary multiobjective optimization of computationally expensive problems. IEEE Transactions on Evolutionary Computation, 23(1), 15-28.

283: Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J. and Sculley, D., 2017, August. Google vizier: A service for black-box optimization. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1487-1495).

284: Gupta, A., Ong, Y. S., & Feng, L. (2015). Multifactorial evolution: toward evolutionary multitasking. IEEE Transactions on Evolutionary Computation, 20(3), 343-357.

285: Gupta, A., Ong, Y. S., Feng, L., & Tan, K. C. (2016). Multiobjective multifactorial optimization in evolutionary multitasking. IEEE transactions on cybernetics, 47(7), 1652-1665.

286: Wei, T., Wang, S., Zhong, J., Liu, D., & Zhang, J. (2021). A Review on Evolutionary Multi-Task Optimization: Trends and Challenges. IEEE Transactions on Evolutionary Computation.

287: Liaw, R. T., & Ting, C. K. (2019, July). Evolutionary manytasking optimization based on symbiosis in biocoenosis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 4295-4303).

288: Tang, J., Chen, Y., Deng, Z., Xiang, Y., & Joy, C. P. (2018, July). A Group-based Approach to Improve Multifactorial Evolutionary Algorithm. In IJCAI (pp. 3870-3876).

289: Gupta, A., Zhou, L., Ong, Y.S., Chen, Z. and Hou, Y., 2022. Half a dozen real-world applications of evolutionary multitasking, and more. *IEEE Computational Intelligence Magazine*, *17*(2), pp.49-66.

290: Zhang, Y., Hoang, T.N., Low, B.K.H. and Kankanhalli, M., 2017. Information-based multi-fidelity Bayesian optimization. In NIPS Workshop on Bayesian Optimization.

291: Shakeri, M., Miahi, E., Gupta, A., & Ong, Y. S. (2022). Scalable Transfer Evolutionary Optimization: Coping with Big Task Instances. IEEE Transactions on Cybernetics.

292: Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

293: Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650).

294: M. S. Veldhuis, S. Ariëns, R. J. F. Ypma, T. Abeel, and C. C. G. Benschop, Explainable artificial intelligence in forensics: Realistic explanations for number of contributor predictions of DNA profiles, Forensic Sci. Int. Genet., vol. 56, p. 102632, 2022, doi: 10.1016/j.fsigen.2021.102632.

295: A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion, vol. 71, pp. 28–37, 2021, doi: 10.1016/j.inffus.2021.01.008.

296: S. Mohseni, N. Zarei, and E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst., vol. 11, no. 3–4, pp. 1–45, 2021, doi: 10.1145/3387166.

297: A. B. Arrieta et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion, vol. 58, no. October 2019, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

298: D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, XAI-Explainable artificial intelligence, Sci. Robot., vol. 4, no. 37, p. 7120, 2019, doi: 10.1126/scirobotics.aay7120.

299: M. Abdar et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion, vol. 76, pp. 243–297, 2021, doi: 10.1016/j.inffus.2021.05.008.

300: P. Madumal, T. Miller, L. Sonenberg, and F. Vetere., Explainable reinforcement learning through a causal lens, Proc. AAAI Conf. Artif. Intell., vol. 34, no. 3, pp. 2493–2500, 2020, doi: 10.1609/aaai.v34i03.5631.

301: Pearl, J.: Causal inference in statistics: An overview. Statistics surveys (2009)

302: Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., Schölkopf, B.: Learning independent causal mechanisms. In: International Conference on Machine Learning (2018)

303: Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)

304: Suter, R., Miladinovic, D., Schölkopf, B., Bauer, S.: Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In: International Conference on Machine Learning (2019)

305: Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., Jordan, M.: A linear nongaussian acyclic model for causal discovery. Journal of Machine Learning Research (2006)

306: Peters, J., Mooij, J.M., Janzing, D., Schölkopf, B.: Causal discovery with continuous additive noise models (2014)

307: Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research (2016)

308: Huang, B., Zhang, K., Lin, Y., Schölkopf, B., Glymour, C.: Generalized score functions for causal discovery. In: ACM SIGKDD International Conference on Knowledge discovery and data mining (2018)

309: Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE (2021)

310: Besserve, M., Mehrjou, A., Sun, R., Schölkopf, B.: Counterfactuals uncover the modular structure of deep generative models. In: International Conference on Learning Representations (2020)

311: Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)

312: Schölkopf, B.: Causality for machine learning. In: Probabilistic and Causal Inference: The Works of Judea Pearl (2022)

313: Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., Lerchner, A.: Towards a definition of disentangled representations. arXiv preprint (2018)

314: Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: Cambridge UniversityPress (2000)

315: Keith, K.A., Jensen, D., O'Connor, B.: Text and causal inference: A review of using text to remove confounding from causal estimates. In: Annual Meeting of the Association for Computational Linguistics (2020)

316: Feder, A., Keith, K.A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M.E., et al.: Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. arXiv preprint (2021)

317: Zeng, X., Li, Y., Zhai, Y., Zhang, Y.: Counterfactual generator: A weaklys upervised method for named entity recognition. In: Conference on Empirical Methods in Natural Language Processing (2020)

318: Wang, Z., Culotta, A.: Robustness to spurious correlations in text classification via automatically generated counterfactuals. In: Association for the Advancement of Artificial Intelligence (2021)

319: Feder, A., Oved, N., Shalit, U., Reichart, R.: Causalm: Causal model explanation through counterfactual language models. Computational Linguistics (2021)

320: Nan, G., Zeng, J., Qiao, R., Guo, Z., Lu, W.: Uncovering main causalities for long-tailed information extraction. In: Conference on Empirical Methods in Natural Language Processing (2021)

321: Bash, Daniil, Yongqiang Cai, Vijila Chellappan, Swee Liang Wong, Xu Yang, Pawan Kumar, Jin Da Tan et al. Multi-Fidelity High-Throughput Optimization of Electrical Conductivity in P3HT-CNT Composites. Advanced Functional Materials 31, no. 36 (2021): 2102606.

322: Ren, Zekun, Felipe Oviedo, Maung Thway, Siyu IP Tian, Yue Wang, Hansong Xue, Jose Dario Perea et al. Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics. Npj Computational Materials 6, no. 1 (2020): 1-9.

323: Karpatne, Anuj, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on knowledge and data engineering 29, no. 10 (2017): 2318-2331.

324: von Rueden, Laura, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch et al. Informed Machine Learning—A Taxonomy and Survey of Integrating Knowledge into Learning Systems. arXiv preprint arXiv:1903.12394 (2019).

325: Karniadakis, George Em, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. Nature Reviews Physics 3, no. 6 (2021): 422-440.

326: Liu, Qiong & Wu, Ying. (2012). Supervised Learning. 10.1007/978-1-4419-1428-6_451.

327: Linus Ericsson, Henry Gouk, Chen Change Loy, Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, Advances and Challenges. (2021). https://arxiv.org/abs/2110.09327

328: Foivos Ntelemis, Yaochu Jin, Spencer A. Thomas. A Generic Self-Supervised Framework of Learning Invariant Discriminative Features. (2022). https://arxiv.org/abs/2202.06914

329: Longlong Jing, Yingli Tian. Self-Supervised Visual Feature Learning with Deep Neural Networks: A Study. (2019). https://arxiv.org/abs/1902.06162

330: Shevlin H, Vold K, Crosby M, Halina M. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. EMBO Rep. 2019 Oct 4;20(10):e49177. doi: 10.15252/embr.201949177. Epub 2019 Sep 18. PMID: 31531926; PMCID: PMC6776890.

331: Jun JJ et al. Fully integrated silicon probes for high-density recording of neural activity. Nature. 2017 Nov;551(7679):232-6.

332: Graves A, Wayne G, Danihelka I. Neural turing machines. arXiv preprint arXiv:1410.5401. 2014 Oct 20.

333: Vaswani A et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

334: Amirhossein Tavanaei, et al. "Deep learning in spiking neural networks." Neural networks 111 (2019): 47-63.

335: Bing Han, et al. "Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

336: Richards BA et al. A deep learning framework for neuroscience. Nature neuroscience. 2019 Nov;22(11):1761-70.

337: Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. Nature. 2019 Apr;568(7753):493-8.

338: Willett FR et al. High-performance brain-to-text communication via handwriting. Nature. 2021 May;593(7858):249-54.

339: Ponce CR et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell. 2019 May 2;177(4):999-1009.

340: Bashivan P, Kar K, DiCarlo JJ. Neural population control via deep image synthesis. Science. 2019 May 3;364(6439):eaav9436.

341: “Superteams: Putting AI in the group”. (2020). https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2020/human-ai-collaboration.html

342: “Huge potential of human - AI collaboration”. (2020). https://tectales.com/ai/huge-potential-of-human-ai-collaboration.html

343: “Procter & Gamble looks to the future with augmented writing”. (2018).

https://textio.com/blog/procter-gamble-looks-to-the-future-with-augmented-writing/13035166567

344: “The Future of Customer Service Is AI-Human Collaboration”. (2019). https://sloanreview.mit.edu/article/the-future-of-customer-service-is-ai-human-collaboration/

345: Jeff Schwartz et al. “Future of work initiatives promise lots of noise and lots of activity, but to what end?”. MIT Sloan Management Review, February 20, 2019.

346: McCaffrey, Tony. (2018). Human-AI Synergy in Creativity and Innovation. 10.5772/intechopen.75310.

347: Mohd Naveed Uddin. (2019). Cognitive science and artificial intelligence: simulating the human mind and its complexity. https://doi.org/10.1049/ccs.2019.0022.

348: Wei Xu, Marvin J. Dainoff, Liezhong Ge & Zaifeng Gao. (2022). Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2022.2041900

349: Petrat, D. Artificial intelligence in human factors and ergonomics: an overview of the current state of research. Discov Artif Intell 1, 3 (2021). https://doi.org/10.1007/s44163-021-00001-5

350: de Lima, E.S., Feijó, B. (2019). Artificial Intelligence in Human-Robot Interaction. In: Ayanoğlu, H., Duarte, E. (eds) Emotional Design in Human-Robot Interaction. Human–Computer Interaction Series. Springer, Cham. https://doi.org/10.1007/978-3-319-96722-6_11

351: Cañas JJ (2022) AI and Ethics When Human Beings Collaborate With AI Agents. Front. Psychol. 13:836650. doi: 10.3389/fpsyg.2022.836650

352: "Social Computing", introduction to Social Computing special edition of the Communications of the ACM, edited by Douglas Schuler, Volume 37, Issue 1 (January 1994), Pages: 28 – 108

353: Kenji Doya, Arisa Ema, Hiroaki Kitano, Masamichi Sakagami, Stuart Russell. Social impact and governance of AI and neurotechnologies. https://arxiv.org/pdf/2112.15459.pdf

354: “Here’s why human-robot collaboration is the future of manufacturing”. (2020). https://www.weforum.org/agenda/2020/08/here-s-how-robots-can-help-us-confront-covid/

355: “Researchers aim to improve human-robot collaboration in industrial workplaces”. (2022). https://www.safetyandhealthmagazine.com/articles/22215-researchers-aim-to-improve-human-robot-collaboration-in-industrial-workplaces

356: Ming-Hui Huang, Roland T. Rust. (2018). Artificial Intelligence in Service. https://doi.org/10.1177/1094670517752459

357: Gruau, F., & Whitley, D. (1993). Adding learning to the cellular development of neural networks: Evolution and the Baldwin effect. Evolutionary computation, 1(3), 213-233

358: Ziyin, L., Li, B., Simon, J. B., & Ueda, M. (2021). SGD May Never Escape Saddle Points. arXiv preprint arXiv:2107.11774.

359: Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. Nature Machine Intelligence, 1(1), 24-35.

360: Fernando, C., Sygnowski, J., Osindero, S., Wang, J., Schaul, T., Teplyashin, D., ... & Rusu, A. (2018, July). Meta-learning by the baldwin effect. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 1313-1320).

361: Cui, X., Zhang, W., Tüske, Z., & Picheny, M. (2018). Evolutionary stochastic gradient descent for optimization of deep neural networks. Advances in neural information processing systems, 31.

362: Salimans, T., Ho, J., Chen, X., Sidor, S., & Sutskever, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864.

363: Liu, G., Zhao, L., Yang, F., Bian, J., Qin, T., Yu, N., & Liu, T. Y. (2019, July). Trust region evolution strategies. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4352-4359).

364: Gangwani, T., & Peng, J. (2018). Policy optimization by genetic distillation. In 6th International Conference on Learning Representations, ICLR 2018.

365: Conti, E., Madhavan, V., Petroski Such, F., Lehman, J., Stanley, K., & Clune, J. (2018). Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. Advances in neural information processing systems, 31.

366: Gruau, F., & Whitley, D. (1993). Adding learning to the cellular development of neural networks: Evolution and the Baldwin effect. Evolutionary computation, 1(3), 213-233.

367: Zhang, N., Gupta, A., Chen, Z., & Ong, Y. S. (2021). Evolutionary machine learning with minions: A case study in feature selection. IEEE Transactions on Evolutionary Computation, 26(1), 130-144.

368: Nomura, M., Watanabe, S., Akimoto, Y., Ozaki, Y., & Onishi, M. (2021, May). Warm starting cma-es for hyperparameter optimization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 10, pp. 9188-9196).

369: Liu, Y., Sun, Y., Xue, B., Zhang, M., Yen, G. G., & Tan, K. C. (2021). A survey on evolutionary neural architecture search. IEEE transactions on neural networks and learning systems.

370: Wang, Y., Xu, C., Qiu, J., Xu, C., & Tao, D. (2018, July). Towards evolutionary compression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2476-2485).

371: Choong, H. X., Ong, Y. S., Gupta, A., & Lim, R. (2022). Jack and Masters of All Trades: One-Pass Learning of a Set of Model Sets from Foundation Models. *arXiv preprint arXiv:2205.00671*.

372: Real, E., Liang, C., So, D., & Le, Q. (2020, November). Automl-zero: Evolving machine learning algorithms from scratch. In International Conference on Machine Learning (pp. 8007-8019). PMLR.

373: Gupta, A., & Ong, Y. S. (2018). Memetic computation: the mainspring of knowledge transfer in a data-driven optimization era (Vol. 21). Springer.

374: Ong, Y. S., Lim, M. H., & Chen, X. (2010). Memetic computation—past, present & future [research frontier]. IEEE Computational Intelligence Magazine, 5(2), 24-31.

375: Duan, J., Yu, S., Tan, H. L., Zhu, H., & Tan, C. (2022). A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence

376: Duan, J., Yu, S., Tan, H. L., & Tan, C. (2020, October). Actionet: An interactive end-to-end platform for task-based data collection and augmentation in 3d environment. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 1566-1570). IEEE.

377: Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., ... & Petersen, S. (2016). DeepMind lab. arXiv preprint arXiv:1612.03801.

378: Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., & Torralba, A. (2018). Virtualhome: Simulating household activities via programs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8494-8502).

379: Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... & Batra, D. (2019). Habitat: A platform for embodied ai research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9339-9347).

380: Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., ... & Su, H. (2020). Sapien: A simulated part-based interactive environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11097-11107).

381: Ecoffet, A., Clune, J., & Lehman, J. (2020, July). Open Questions in Creating Safe Open-ended AI: Tensions Between Control and Creativity. In ALIFE 2020: The 2020 Conference on Artificial Life (pp. 27-35). MIT Press.

382: “National Supercomputing Centre”. https://www.nscc.sg/

383: “A sneak peak of Singapore’s latest petascale supercomputer”. (2022). https://www.nscc.sg/wp-content/uploads/2022/07/NewsBytesFULL-Jul2022-issueFinal.pdf

384: “Supercomputer Spotlight: ASPIRE 2A to arrive in 2022”. (2022). https://www.asianscientist.com/2022/03/print/supercomputer-spotlight-aspire-2a-to-arrive-in-2022/

385: Nossokoff, M., Riddle, M., Norton, A., Sorensen, T., Conway, S., & Joseph, E. (2021). 2021 Strategic Opportunity Assessment and Study of the HPC Roadmap in Singapore. Hyperion Research.

386: Sarma, Sankar & Deng, Dong-Ling & Duan, Lu-Ming. (2019). Machine learning meets quantum physics. Physics Today. 72. 48-54. 10.1063/PT.3.4164.

387: Prashant, S. (2005). A Study on the basics of Quantum Computing. https://doi.org/10.48550/arXiv.quant-ph/0511061

388: C. Ciliberto et al., "Quantum machine learning: a classical perspective," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 474, no. 2209, p. 20170551, 2018.

389: Chowdhury, M. (24 March, 2022). How Can Quantum Computing Change the World? Retrieved from Analytics Insight: https://www.analyticsinsight.net/how-can-quantum-computing-change-the-world/

390: I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," Nature Physics, vol. 15, no. 12, pp. 1273-1278, 2019.

391: J. Gibbs et al., "Dynamical simulation via quantum machine learning with provable generalization," arXiv preprint arXiv:2204.10269, 2022.

392: H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill, "Provably efficient machine learning for quantum many-body problems," arXiv preprint arXiv:2106.12627, 2021.

393: G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," Science, vol. 355, no. 6325, pp. 602-606, 2017, doi: doi:10.1126/science.aag2302.

394: J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," Nature Physics, vol. 13, no. 5, pp. 431-434, 2017.

LIST OF ACRONYMS

A*STAR: Agency for Science, Technology and Research

AGI: Artificial General Intelligence

AI: Artificial Intelligence

AIGC: Artificial Intelligence Generated Content

AISG: Artificial Intelligence Singapore

ANN: Artificial Neural Network

AutoML: Automated Machine Learning

BAAI: Beijing Academy of Artificial Intelligence

BCI: Brain-Computer Interface

BO: Bayesian Optimization

CC: Quantum-Inspired Machine Learning

CF: Catastrophic Forgetting

CFAR: Centre for Frontier AI Research

CI: Computational Intelligence

CL: Continual Learning

CNN: Convolutional Neural Network

CQ: Classical-Quantum Machine Learning

CSIRO: Commonwealth Scientific and Industrial Research Organization

CV: Computer Vision

DAI: Design and Artificial Intelligence

DBS: Development Bank of Singapore

DNN: Deep Neural Network

EA: Evolutionary Algorithm

EPFL: École Polytechnique Fédérale de Lausanne

ESG: Environmental, Social and Governance

ETHZ: Eidgenössische Technische Hochschule Zürich

FICO: Fair Isaac Corporation

FL: Federated Learning

FRC: Foundation Research Capabilities

FWCI: Field Weighted Citation Index

GAN: Generative Adversarial Network

GCN: Graph Convolutional Network

GDPR: General Data Protection Regulation

GNN: Graph Neural Network

GP: Gaussian Process

GPT-3: Generative Pre-trained Transformer 3

GPU: Graphics Processing Unit

HAIC: Human-AI Collaboration

HPC: High Performance Computing

IoT: Internet of Things

LLM: Large Language Model

LSTM: Long Short-Term Memory

MAML: Model-Agnostic Meta-Learning

MAS: Multi-Agent Systems

MeL: Meta Learning

MFEA: Multifactorial Evolutionary Algorithm

ML: Machine Learning

MOH: Ministry of Health

MTL: Multi-Task Learning

NGO: Non-Governmental Organization

NSCC: National Supercomputing Centre

NLP: Natural Language Processing

NRF: National Research Foundation

NTU: Nanyang Technological University

NUS: National University of Singapore

OoD: Out of Distribution

PDPA: Personal Data Protection Act

QC: Quantum Computer

QC ML: Machine Learning Aided Quantum Computing

QFT: Quantum Fourier Transform

QKD: Quantum Key Distribution

QML: Quantum Machine Learning

QPE: Quantum Phase Estimation

QQ: Purely Quantum Machine Learning

R&D: Research and Development

RIE2025: Research, Innovation and Enterprise 2025

RL: Reinforcement Learning

SC: Supercomputer

SDG: Social Development Goals

SGD: Stochastic Gradient Descent

SL: Supervised Learning

SMU: Singapore Management University

SNN: Spiking Neural Network

SSL: Self-Supervised Learning

STEM: Science, Technology, Engineering and Mathematics

SUTD: Singapore University of Technology and Design

SVM: Support Vector Machine

TinyML: Tiny Machine Learning

TL: Transfer Learning

TPU: Tensor Processing Unit

UQ: Uncertainty Quantification

WEF: World Economic Forum

XAI: eXplainable Artificial Intelligence

Commissioned by

NATIONAL RESEARCH FOUNDATION
PRIME MINISTER'S OFFICE
SINGAPORE