DeepScience

Cognitive Artificial Intelligence for Health and Climate

Deep Models, Interpretability, and Decision Support

Samit Shivadekar

Cognitive Artificial Intelligence for Health and Climate: Deep Models, Interpretability, and Decision Support

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

DeepScience

Published, marketed, and distributed by:

Deep Science Publishing, 2025
USA | UK | India | Turkey
Reg. No. MH-33-0523625
www.deepscienceresearch.com
editor@deepscienceresearch.com
WhatsApp: +91 7977171947

ISBN: 978-93-7185-581-5

E-ISBN: 978-93-7185-745-1

https://doi.org/10.70593/978-93-7185-745-1

Citation: Shivadekar, S. (2025). *Cognitive Artificial Intelligence for Health and Climate: Deep Models, Interpretability, and Decision Support*. Deep Science Publishing. https://doi.org/10.70593/978-93-7185-745-1

Preface

The intersection of health, climate, and artificial intelligence represents both a challenge and an opportunity. This book explores how Cognitive AI — combining deep learning, reasoning, and interpretability — can address pressing issues in healthcare and climate science while supporting trustworthy, informed decision-making.

Focusing on deep models, explainable AI, and decision-support frameworks, we examine methods that go beyond automation to enable human–AI collaboration in complex, high-stakes environments. Through research insights and real-world case studies, the book bridges theory and practice for researchers, practitioners, and decision-makers seeking resilient and transparent AI solutions.

It is my hope that this work inspires the development of AI systems that are as trustworthy as they are innovative, serving society's urgent needs with clarity, fairness, and resilience.

In writing this book, I have drawn on my own experiences in research, teaching, and collaboration across disciplines, as well as the invaluable contributions of the AI, health, and climate research communities. I am grateful to colleagues, students, and partners whose insights have shaped the ideas presented here. May this work inspire continued exploration into AI systems that serve humanity's most urgent needs — and do so with clarity, fairness, and resilience.

SAMIT SHIVADEKAR

Table of Contents

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 1: Cognitive Artificial Intelligence for Societal Resilience

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. The Convergence of Health and Climate Challenges

In 1972, an early warning was made against the dangers of the interplay of three main malfunctioning factors in global system failure: pollution, overconsumption, and inequity. These problems were structuring elements in exponential growth with changing nature of their interplay across time. Fifty years later, we are confronting global consequences of climate change and the COVID-19 pandemic that are both consequences of a mismanaged global system [1-3]. Systems of global connectivity can generate a massive convenience for everyone, promoting access to most material resources and communication range. But the same systems are generating inequity, fragility, exploitation of finite resources, and degradation of leadership, trust, and collaborative intentions.

Population and economic growth, together with biophysical changes feeding back into the economy, mean that greenhouse gas emissions will need to peak and decline within the next decade to avoid the worst of climate change, and global poverty needs to be eradicated while reducing biodiversity loss within the next decade [2,4]. The technology available to simultaneously achieve these ambitious goals is not scarce: renewables already supply the largest fraction of gigawatts, batteries are becoming cheaper and allowing the decarbonization of transport. But the organizations capable of mobilizing and coordinating ecosystems, value chains, decisions, and investments at the required scale and speed are few, scarce, and fragile. These groups need to be supported by flexibility of government policy, access to talent and innovation, and skills training and human capital investment. Without these systemic enablers, temporary spurts of climate-tech decarbonization or green finance, for example, run the risk of being just that: short-lived spurts [5-8].

2. Defining Cognitive AI: Symbolic, Neural, and Hybrid Approaches

Cognitive Artificial Intelligence (Cognitive AI) is an emerging field of research at the intersection of Artificial Intelligence (AI) -- specifically traditional AI, or Symbolic AI, and Machine Learning, specifically Representation Learning, notably Neural Networks and more generally Deep Learning applied to the problem of representation learning or acquiring features from experience. Cognitive AI builds on the foundations of Symbolic AI [6,9]. Knowledge Representation and Reasoning is where Cognitive AI differs from Symbolic AI. Symbolic AI emphasizes the representation of symbolic systems, whereas Cognitive AI emphasizes additional cognitive principles such as experience and learning--both cognitive learning, which is inspired by cognitive science and developmental psychology. Cognitive AI builds on the foundations of Symbolic AI, symbolic cognitive architecture, found its origins in the field of Computer Science, developed by various researchers. More generally, Cognitive AI focuses on how to make computers more intelligent in the very specific and general way that as humans we can learn different types of skills, conceptual knowledge, perceptual knowledge, social-cognitive knowledge, and so on, from a combination of embodied experience -- the automatic supervised embedding style of learning based on physical presence and agent-environment interaction -- with the tools of Language, Imagination, and Knowledge [10-12].

Over the years, several attempts at Cognitive Architectures have proposed an initial solution to this worthy endeavor but providing increasingly more expressive systems. The more recent works on Neural-symbolic Models include Neural Agent-environment Interaction, Symbolic Program Generation, Code Generation, Visual Gen-modalities like language, code, and even other visual Developing Visual Models of the World, Knowledge-based Learning of Neural Network Architectures -- Range of Componential Cognitive Skills such as Analogy, Induction and Abduction, Reasoning about Agents and Intention Recognition, Causo-matic. With the very recent developments in Foundation Models for multimodal agents and social robots, we appear to be on the verge of something new, combining very efficient experience driven Statistically based Learning, Cognitive Development and Plasticity.

3. Strategic AI for Decision-Critical Environments

AI supports decision making in many different areas. Commercial, marketing, business operations, production and logistics decisions are facilitated by operatively oriented AI tools [7,13-16]. Numerous consumer-oriented AI applications have become available. However, none of the publicly available AI tools is specifically designed to facilitate the decision making in the societal and environment area where action have high, long term implications and are essentially irreversible. Such applications would be used for

building long-term, sustainable, resilient societies and environments. Rather, their area of application is more narrow, focusing on preventing and reacting to strong shocks and crises. They lack the stringent quality assurance required in the societal context, and they address one of the many aspects of complexity and uncertainty encountered in this context [2,17-19]. They are designed to be used by practitioners like administrators or security experts and propagate results which can directly be converted into operational action.

In this paper, we introduce strategic AI, i.e. AI designed for supporting strategic decision making in complex, dynamic decision-critical environments. We outline requirements and criteria on how to define such AI applications, their structure and functionalities. We discuss how to build them, with which technologies and methods they can be implemented, the specific knowledge base requirements and their verification. And how quality assurance can be implemented and discussed design settings which consider the fact that such strategic decisions are usually made in group processes by human decision makers, furthermore, stressing the issue of societal validity of the group recommendations put forward by these AI applications.

4. The Role of AI in Public Health

Many entities use AI tools extensively to support various tasks to bring effective solutions for many diverse public health challenges [3,20-23]. These included traditionally academic institutions and research consortia, commercial entities, and many private research organizations and consortia, and government agencies and NGOs. The roles of AI in public health broadly relate to these functions: bio surveillance and epidemic intelligence, situational awareness and early warning, Explainable AI, pandemic modeling, diagnosis of infectious diseases, prediction of disease behavior, vector surveillance and insect control, vaccine development and manufacturing, AI-enabled clinical workflows, smart hospitals, hospital decision support systems, risk stratification and clinical management, improvements in community health, mental health and wellness, online mental health support, public engagement, communications and mHealth, AI in global health, AI in furthering the SDGs, public private partnerships for equitable AI-enabled health solutions, ethical AI in public health.

Bio surveillance is the monitoring of indicators of disease, injuries, and/or death through different syndromes. It is an early warning system that facilitates the detection of potential outbreaks before they occur and thereby potentially enables their earlier containment and control [9,24-26]. Used effectively, bio surveillance could allow to identify disease hotspots and track the changes in the spread of disease and health risks over time. AI allows taping of the exponential increase in adoption of smartphones,

internet use, and social media to identify diseases, their predictors and symptoms, and geolocate them to monitor, visualize, and forecast infectious disease events—with a goal of potential rapid incorporation of data into public health action.

5. Climate Change and AI Solutions

The world is being reshaped by the challenges of climate change. From fossil fuel carbon emissions to deforestation, urbanization, over-consuming, and waste-producing societies are reshaping planetary systems, moving the Earth into unknown states with unpredictable local and global consequences, near and far. Many of these consequences will be harmful and, especially where the developing world is concerned, dire and concerning. At present more than half of humanity lives in an urbanized state and this figure continues to grow [27-29]. Urban populations are stressed by heat, drought, floods, and rain and face unforeseen challenges accompanied by migrations and disasters. Disconnection between nature and city systems induces stream damage, durability, infrastructure degradation, and increased instability. All these threats require an immediate response.

Debates about the role of Artificial Intelligence in Climate Solutions are rife, and reports abound. But what is remarkable are the facts-based AI Use Cases, such as goals applied to Climate AI Engines and Machines that have been compiled and freely shared. Believe it or not, while this may represent only a modest fraction of real-world potential AI applications, already more than 650 are collecting. These applications are all the more important because the potential for impact is immense. So much so that, if done correctly, Climate Change could be the greatest accelerant for addressing inequality and other development goals that we have seen within the next fifty years. To ensure the success of the AI movement, it is important to remember that it is not all about technology. Lifescape 2.0, or Intelligent Sustainable Landscapes, will engage appropriate AI in synergy with Urban Design, Architecture, Infrastructure, Nature Conservation, and Urban Planning – but at scale and with proper funding mechanisms in place.

6. Ethical Considerations in Cognitive AI

We consider how can we think responsibly about Cognitive AI and its potential societal effects. We explore some concrete ethical principles that may serve as guides: beneficial AI, high-level human control over AI, Transparency and Explainability, Broader access to benefits, Privacy, Societal agency, and Mitigation of global catastrophic risk [30-32]. Cognitive work can lack transparency and reconciliation is necessary. Building even capable systems cannot be done in isolation. Their orchestration requires deep collaboration. The models and simulations, inside biomedical science for example, can raise hard ethical questions: who gets to decide?

For which purposes? About what abilities? Such negotiations require careful reflection on the values invested into decision-making. Broad and publicly accessible models require different consideration compared to those fraying in deep technical proximity among expert peers. This is especially true for ethical boundaries. Both regulators and AI developers need to find ways to share knowledge and concepts in order to agree on what constitutes proper usage. Enacting and regulating fair usage is a challenge we want to address, even if tentatively, when proposing experiential education grounded natural-language ethical with decision-making pathways. These pathways enable tracking, accountability and help permit wide engagement, ethically grounded societal reconciliation.

7. Data Privacy and Security in AI Applications

Data privacy is a crucial concern when developing AI for societal resilience, especially in areas such as sharing medical, financial, and communications data. Regulation, consent, identity verification, and automation of data governance mechanisms can help maintain privacy online while ensuring data-driven machine learning [9,33-35]. Privacy and security-by-design principles can be effectively implemented in ML models in terms of data control, obfuscation, and leakage prevention. Methods such as creating federated learning pipelines, differential privacy, and cryptographic protocols can offer solid guarantees. Regulation and public-private partnerships can help strengthen these efforts, particularly in contexts of critical socioeconomic risks.

AI and ML applications have critical security and privacy implications across sectors. For instance, in safety-critical scenarios, such as autonomous vehicles or healthcare systems, AI and ML systems need to satisfy additional verification and validation methods to ensure that data integrity, data availability, and data confidentiality are safeguarded [36-39]. Security and privacy concerns permeate the AI stack, meaning from the data to the computations and finally the models. Various attack vectors can lead to significant implications on the functioning of AI systems, or outright make these nonfunctioning, for instance, implementing data poisoning or backdoor attacks. Regulations can help clarify the guidelines and harm horizon for threat modeling in the design of AI systems to avoid catastrophic threats to humans or infrastructure.

8. AI for Disaster Response and Management

Disasters, both natural and anthropogenic, can disrupt individuals, societies, the economies and the ecology across any region in the world. Global recent events have demonstrated the unpredictability and threat of disasters, but also highlighted the importance of human resilience against such hazards. AI-based solutions are being increasingly adopted for efficient disaster response and management plans [6,9].

Recent years have also witnessed collaborative efforts that are using AI to develop better predictive, mitigating, response and recovery capabilities against disasters.

As disasters can be classified as either natural or fabricated by humans, AI has been applied to numerous sub-tasks of disaster management driving efforts in situational awareness, data insights, response planning or recovery. Natural disasters such as floods, wildfires, cyclones, earthquakes, etc., threaten life and property across the globe regularly and thus, AI for disaster response has focused on actionable insights to reduce the impact. AI for disaster management has, therefore, relied heavily on data-driven solutions using satellite and drone-based images, social media posts, and other geographic information system data. Orchestrating substantial advancements in expert systems, AI and related emerging technologies can work together to create synergies that are necessary for informed decision making during disasters. Human behaviours during disasters have also been simulated using gaming platforms enhanced by ML methods to predict the help-seeking behaviour, which is useful for efficient disaster management.

9. Interdisciplinary Approaches to AI Implementation

Cognitive AI projects are undertaken by people from a variety of domains, including humanities, arts, sciences and technology, social and political sciences, management, and industry. Some of them are hired as AI experts by global tech companies, while others use AI algorithms as research tools. To unpack the different roles that different knowledge domains play, and on what basis AI experts agree to combine as HAI (Human-AI) team, we propose an interdisciplinary taxonomy.

What we call critical AIs are those used to investigate cognition and intelligence, which usually take as their test-beds AI systems capable of doing or deciding things that are interesting from a cognitive, ethical, philosophical, or even legal point of view. A well-known example is system Fairness – which allows sensitive data to be omitted from the training set, thus protecting citizens' privacy. Natural language processing (NLP), and in particular algorithm bias, is one of the most heavily examined AI domains in this respect, since AI projects for developing NLP functionalities for different languages have been publicly denounced to be biased [2,6]. A first line of criticism states that these language technologies have been developed without representing the diversity of the world's languages – only English, due to its overwhelming presence on the internet, is being represented, while 6,000 languages are today spoken on earth. A second issue relates to how well NLP pipelines are capable of handling multi-linguistic requirements. Solutions for low-resource languages, and for dialects of Asia and Africa, for which there are no word embeddings and translation systems publicly available, are still to be developed.

10. Case Studies of Cognitive AI in Action

Cognitive AI is an evolving field pioneered by a growing number of organizations and initiatives that use the concepts of cognitive AI. We have documented the use cases of a selection of those organizations in the judgment that they have been the forerunners most relevant for demonstrable capability development and skill discovery for societal resilience. These use cases also provide rescue packages and funding alternatives for post-pandemic recovery, for companies reliant on tourism and hard-hit sectors, to creatively help people overcome the unwelcome consequences of isolation, loneliness, and depression. It is hoped the collaborative tools, practice, capabilities, and networks built from these programs can help make societies resilient, build their capability and toolkit to deal with mass disruptions using the virtual environment and collective intelligence in convergence with the real world to build actionable knowledge and productive policy.

This research included a meta-study of at least ten initiatives that are cognitive AI-enabled, and not simply custodians of cognitively augmented human capital. The collaborative practice research has different use modes for mobilising modal curiosity, roles of players from using mood, touch, and feeling in forms of interaction. Exploring social models allows agents to find desired or desired potential dynamics for co-creation, decision tool availability in different application flows, graduation of capability through agency and commitment all have implications for the design of intelligent systems. Case studies of existing operations illustrate potential avenues available to operations designers wanting to build, or expand on existing operations, employing cognitive AI.

11. Challenges in Developing Cognitive AI Systems

Cognitive AI systems are intelligent systems capable of performing higher-level tasks traditionally associated with human intelligence. The consultation provokes a series of very serious questions about capabilities, intrinsic motivation, values, and goals. The answers to these questions can also indicate that we are still far from being able to develop AI truly capable of cognitive work. Unlike traditional AI systems, which largely imitate the status of task and situational representations without inferring these representations, cognitive AI utilizes its own internal symbolic representations and models of the physical and social world to carry out inference and reasoning. The potential emergence of a type of cognitive AI, capable of autonomously developing the symbolic representations and models on which dedicated AI systems operate, offers a non-trivial meta-solution to the long-standing problem of the harsh requirement for knowledge engineering in traditional AI. The question however, we ask in this essay is how cognitive AI can positively contribute to societal resilience.

Highly specialized algorithms are no longer able to tackle the challenges that our society is continuously facing. Usually these problems share a similar nature, are complex composed by multiple recursive problems that need to be accounted for in a parallel manner simultaneously during their treatments. A possible solution might be the develop of cognitive informa-AI systems that can autonomously learn from being in touch with expert users and test many different solutions at the same time. Some initial steps in this direction have already been explored by promoting autonomous cognitive architecture that integrate reservoir computing with multi-agent artificial ecosystem, which instantiate echo chambers composed by AI systems alternative solutions in parallel to the one proposed by human expert users. However, research on this topic is still at its infancy and many challenges need to be faced.

12. Future Trends in Cognitive AI

The future design landscape will have such devices as eye glasses that will detect automatically when the user should improve their attention outside of them, or for attention lowering when it becomes harmful. Such devices will have embedded Cognitive AI functions with high throughput, low energy consumption, embedded information reprogramming, communications, and trust. They also will be embedded in large, universal, and deployable societal or individual Artificial Humanoid Intelligent characters [7,13,16]. Such Humanoids would serve as role-models for teaching purposes, motivation, local companion in services of everyday life as well as reprisal of the majority of tasks done by contemporary low intelligence robots or applications. Universal group of Humanoids of such intelligence and reliability with allocated sociogenetic humor would also participate in planning by codesigning and steering by reprogramming human cognitive joint activity during all moments of real solving of complex goals. Such design mission can be understood as delegation to AI-amending system of subgoals allocation, task stages organization, sharing of research, education, and motivational resources, collection and estimation of big dataflow results for Earth feedback, responsibility, trust, control.

The last part of the imagination design stream will be architecture-genetic group and its collective imitation and collective responsibility for the digital twins and Humanoids for everyone's astrobiogenetic goals for case of closing, power-consuming road for the Earth sustainable development in concordance with naturality laws. Interactive Emotional Humanoids, maintaining user’s awareness of instant local goals and long-term horizon objectives program, would allow to achieve such a concord with fast rehabilitation. Collective image about how the development of human communities allows solving of these tasks will create the best assemblages of Cognitive Assistants, and Chairs as the ultimate goals of the digital societies. Creating of such digital environment would be role of the best design and cognitive science of

the communities of scientists, educators, artists, specialists who grew in discussed cognitive and ethical horizon.

13. Collaborative AI and Community Engagement

Collaborative AI makes AI technologies and tools available and reliable to a broad group of stakeholders, preferably those with the specific domain expertise needed to identify appropriate solutions, dynamically modify the processes as needed, and validate and take responsibility for the outcomes of the cognitive technology-driven processes. AI solutions developed by domain experts are aligned with domain knowledge and community-owned institutional knowledge and are better integrated into existing workflows. Collaborative AI also enables more users to leverage the value of AI technologies at any level of the investment and expertise.

There is an assortment of technologies and processes that fall under CAI, addressing a spectrum of disciplines, knowledge domains, and methods ranging from original AI research in knowledge representation and reasoning to integrated development environments that assist users in specific tasks often supplemented by easy-to-use rule-based systems. The aims of these tools usually include the democratization of technology, addressing the knowledge gap requiring AI expertise, increasing the productivity and accuracy of AI work, and supplementing any limitations in the initial and continuing training of the AI models with user-provided expertise, constraints, and/or feedback. This is a critical necessity for addressing the multidisciplinary and multi-stakeholder problems associated with climate resiliency, as well as a range of additional application domains.

14. AI for Sustainable Development Goals

Artificial intelligence (AI) holds promise in advancing global sustainable development agendas. It can help track the progress of various Sustainable Development Goals (SDGs), monitor trends in long-term development, improve information collection efficiency, enhance the capacity to glean insight from data analytics, provide innovative tools to deal with new issues in sustainable development, create changes in behavior that redirect trajectories towards sustainability, and improve the efficiency of human and natural resource utilization. AI algorithms can assist in estimating and tracking a large number of Vital Registration Systems (VRS) that record live births or deaths from a distance. Tools such as natural language processing can be used to process different types of data for flexible coding of specific Geo-Referenced Development Concepts (GRDC) for the entire world due to the little availability of signals from governments that actually code it for the VRS.

Major AI techniques can also address 8 out of 17 SDGs including poverty, health, education, water, inequality, urban research, and development partnerships. There are many applications of AI in achieving the 2030 Agenda for Sustainable Development, which will be explicitly talked about in this chapter using examples from around the world. AI has already been used to assist more than 4000 indigenous communities around the world to mitigate damage from droughts or floods while helping them better adapt and avert damage from climate change. Countries as diverse as Singapore, Madagascar, Israel, and Kazakhstan have developed or proposed AI solutions for sustainable development. Many countries and regions have all touted AI to accelerate the achievement of the SDGs. There are many positive potential applications highlighted in the literature, but there are also as many if not more darker or negative cases.

15. Building Resilient AI Frameworks

Cognitive or Understanding AI systems, as we have described them throughout, are as good as the structure and content of their working, long, or world model. The architecture should be modular and tap only content and aspects of reality that are relevant in the moment for how the system is to act. The content in turn has to be derived from a grounded, embodied, and multi-sourced approach that centers on the lived experience of the organisms that have to actually interact with the real world. Moreover, content organization has to facilitate efficient learning and recall, through complex hierarchies, latent semantic indexing concepts, and Nominalization Clusters ensuring the differentiated activation of distinct modality abstractions within which realization sequences can cascade. The aim of these modules is to directly influence the process of how behavioral commands are generated and what their moment-to-moment content should be, i.e. the what ad how of cognition. This design feature allows for the unique Quality of Guidance of Cognitive AI systems, the moment in a sequence that is actually being worked on in the moment of decision, enabling a massive acceleration and optimization of Cognition-to-Action processing moment-to-moment, compared to an alternative implementation that resorts to priming of associative memories to dictate, via a general mechanism, the direction and probabilistic influence of action from cognition indirectly, in a shot-gun manner [2,4]. Nothing but guiding commands can actually operate on the system's built-in action executors, whose workings still take place as body-based motor and perceptual-motor control loops. So, it is through the moment-to-moment work on a dynamic hierarchy of behavioral commands that the Cognitive AI forms a direct interface with human brains and behavior.

16. Policy Frameworks for AI Governance

Societal resilience perspectives for or against the use of technology are unlikely to temper or mitigate human inhumanity. It would be collusive for any framework to argue the paradox that we need AI to govern society. Policy frameworks for governing AI must acknowledge the propensity for abuse of power that AI brings with it. They should not be only either proactive, in the sense of enabling or providing incentives for innovation, or negative, in the sense of constraining actors and AI technologies. If appropriate agentic and societal capacities are created or exist, it should be possible for both pro- and negative policies to co-occur. Policy frameworks must also acknowledge that actors who are entrenched in societal inhumanity would likely not be interested in cooperating or following even the most positive and incentive-driven policies. Policy frameworks would need to create agentic and societal capacities at multiple levels of governance.

Genealogically, different layers of society may have differing capacities to absorb shocks and crises, and thus would need to be at different levels on the resilience spectrum. Different layers of society are tasked with different scales of compressive powers, and cannot always insulate the more local layers of society from inhumane activities. Just as the field of crisis studies has difficulty in defining what constitutes a crisis, so too morass an area of contention and social constructionism is the area of policy frameworks for the governance of artificial intelligence. At a fundamental level, policy frameworks seek to provide individuals and groups in society with the necessary infrastructures to either band together and become more resilient, or to act become a resilient monad in an increasingly globalized world.

17. The Impact of AI on Employment and Workforce

The impact of AI on workforce is complex, variable, and context-sensitive. Jobs differ in terms of activities performed, skill level, and AI-infection potential. Jobs are displaced and modified, but not created equally. Those exacerbated and stemmed inequalities differ according to country, sector, and sub-groups of workers. Several jobs are likely to be modified, due more of optimization and incremental efficiency explorations rather than a landmark shift and new opportunities driven by groundbreaking breakthroughs. The share of jobs being displaced too show very uneven patterns, lopsided towards low-paid and mid-income classes, hallowing the so-called middle class. However, there are new opportunities in caregiving activities, social services, technological development and maintenance, as well as on eco-sustainability. The path creation process will be of paramount importance for governments, academic institutions, and other stakeholders. A painful period of transition and turmoil is expected.

Governments' policies matter. AI technologies will also disrupt the labor, capital, and goods markets, increasing precariousness, inequalities, and polarizations in wealth distribution. On the labor market side, traditional policies like unemployment insurance financing automation-affected job-seekers' retraining and re-skilling will be likely unsatisfactory in case of fast tech-pacing. Stimulating labor demand via tax reductions or credits, support to R&D, newly created jobs subsidies, and direct public investment on or with private firms, moreover, should be enlarged and/or implemented. Education inequality limits social mobility opportunities, especially for youths from disadvantaged backgrounds. Additionally, societies' attitudes towards entrepreneurship, innovation, and risk-taking play a main role. Tackling labor market asymmetries and imbalances requires time. Negative expected outcomes on labor conditions and quality may have spillover effects on social stability, macro-demand, and political choices, accentuating dissatisfaction and anger responses toward policy-makers and institutions and/or populistic movements.

18. AI and Mental Health Interventions

There is a great push at present to redefine mental health treatment paradigms. In addition to providing counselling, support and therapies, there is increasing interest in acting earlier and using prevention (including self-prevention) as a focus. Since humans spend most of their time alone, intervention through, or use of, technology seems a natural direction for these efforts. AI can augment human capacity by providing faster, more effective and scalable alternatives. These interventions can act at the level of the individual, where AI can track symptoms and moods, even invisibly through wearable or connected devices. When patients do seek treatment, for example in the consultation room or over telemedicine, AI can assist physicians with diagnosis, therapy suggestions, and progression tracking. AI-based software can help during stand-alone interventions. These efforts can range from reinforcement learning capstone therapies to chatbots that provide easy and immediate understanding, tracking and support of low-intensity programs. At a wider level, AI can search and mine data over the population, spotting emerging trends and micro-trends invisible on the surface that can trigger public policy or research investigations.

Embodied AI or robot companions are being increasingly used in specific groups such as children with autistic spectrum disorders or the elderly in hospitals. They act in tandem with a therapist, but are present in between sessions, reducing repetition bandwidth (though some warning of over-reliance is given). Speech understanding provides an intuitive interface for non-expert end-users and is now accurate enough to become a reliable means of interaction. A large number of user interaction studies have been carried out, in which users simply talk to AI agents in natural language, looking for any emotional connection or reflection behavior. These AIs are aligned to represent

feelings, either mimicking or modulating the user's speech, and to offer accurate cues to important subtexts or conversational nodes.

19. Integrating AI into Educational Systems

AI's sweeping capability will affect all dimensions of society, elevating the question as to how AI should be integrated into current and future operations. Education and training are key elements of any considered answer. Education is fundamental both to leverage AI for the good of society and to prepare future generations for functioning in a world pervaded by AI systems as collaborators, assistants, and colleagues. Creative critical thinking, empathy, and emotional and creative intelligence will become the key traits of new work profiles indeed AI will not be able to replicate. Hence, educational models will have to adapt rapidly to integrate AI's strengths and weaknesses in their designs. The democratization of AI exploration is increasing its popularization, fueled by the availability of interfaces that simplify its utilization. However, an understanding of the workings of AI technologies is not universal yet, while a specific knowledge of AI principles remains confined to a narrow group of specialists.

How do we provide continuous education for the population? What is the new function of educational institutions? A faster integration of AI into dedicated curricula requires a massive push from educational systems. What will be the capacity of universities to integrate AI in new professional profiles? How do we formulate regulations to avoid that any answers to these questions will result in widening the gap between digital citizens and digital illiterates? In other terms, how can we elaborate educational policies that both leverage cognitive automation in favor of society and at the same time educate the same society to a metaphorical coexistence with intelligent machines? How will the effects of these policies foreseeably impact future generations? Without a doubt, the reshaping of AI systems will be central to formulating specific answers to these challenging questions that will shape the futures of at least the next 100 years.

20. Public Perception of AI Technologies

Numerous empirical studies explored peoples' views of AI technologies and reflected them in reports published by various organizations. Some of the factors influencing the International Public Perception and Acceptance of AI technologies are knowledge of and interactions with AI technologies, country demographics, culture, inventiveness, trustworthiness, and optimism. The primary area of focus of surveys regarding the International Public Perception of AI is on the need for and concern about how/whether AI technologies will be used; for example, concerns about loss of job opportunities in the automation cycle due to the use of AI technologies. Misinformation about the capabilities of AI technologies may have biased participant

responses. Other areas that surveys have focused on include trust in AI's decision-making and transparency in the AI technology.

Public concerns about AI technologies mostly relate to those domains in which they fear AI technologies may have a direct impact; for example, military applications for weaponized AI. When assessing the benefits of AI utilization, replies were found to differ between general benefits and applications that directly impact the participant's daily life. People were optimistic about the general nature of AI technologies but had negative sentiments as to their own future experiences and thus were cautious and ambivalent about utilizing AI platforms. A recommendation is that those involved in the design of intelligent systems be cognizant of the social impact that their systems may produce and discuss both the negative and positive influences during interaction with users.

The area of Social Robotics connects AI technologies with human interaction to enable robots and robotic systems design for social acceptance. Ethical considerations, and questions about how social robots should be implemented, are areas of concern in which recommendations have been provided; for example, how humans should and should not specifically relate to social robots.

21. Funding and Investment in AI Research

After the boom in AI during the mid1960s–1970s, funding was significantly reduced. Furthermore, during the 1990s and 2000s most government initiatives focused on narrow AI systems with limited societal applications, which increased to some extent during the 2010s. While institutional support for AI research in this narrow domain was strong, an over-reliance on fragile tech startups has characterized the funding structure for robust and trustworthy AI systems. The latter have predominated in pooled and private capital structures that are meanwhile ineffectively coordinated and regulated. Whereas a small circle of number-crunchers have concluded that in the next decade the global economic contribution from developments in global machine learning technologies would reach only $6 trillion and that Northern American production and manufacturing would only gain $2 trillion; not enough has been invested in AI for the sciences and the 50% of the human population not living in Northern America, Western, and Northern Europe. This has limited the global perspective of, and inclusivity towards, unlike any other technology innovation currently, developing a completely new kind of AI that is able to collaborate with humans and address the global challenges as set out by the Sustainable Development Goals.

London has been for some time the world leader in AI investment, especially in the Edge AI and Natural Language Processing domains. Since 2000, it has attracted nearly

one-third of total investment in European AI tech companies. China has been a close follower, successfully replicating and internationally distributing tech ideas for a while. The US was previously the pioneer, but its start-up economy in this area then de-adopted long-term project funding in favor of pure technical ideas. However, more recently the US has regained its first place considering total numbers and combined skills of scientific personnel in almost every AI subdomain, even compared to weekly attractive communities in the Big Quad Countries and to some nascent countries.

22. AI in Urban Planning and Development

Urban areas are inherently centers of economic growth and innovation, while at the same time, they become prone to various stresses, resulting from the accumulation of economic, energy, and environmental-related faults. City systems are becoming ever more complex, as they increasingly rely on technological innovations, especially information and communication technologies that support the constantly increasing exchange of information of the wide variety of actors involved in urban resilience. Considering the broad spectrum of interactions and exchanges among a city's many components and between the city and the surroundings, our main proposition focuses on the synergies that emerge using AI to model the interactions, detect relevant patterns, run forecasts and prescriptive models. These might help city leaders and planners catalyze the support and involvement of the relevant city stakeholders through specific policies or toolkits and playing smart stimuli and aid roles to mitigate the stresses that impact the city resilience and adaptive capacity. Knowing that the capabilities of cities, especially of small and mid-size, are usually constrained when dealing with stresses and shocks, researchers, tool developers and professional planners should find ways to provide cities with reliable and relatively inexpensive AI-based service products. Decision makers may be reluctant to trust the system's output, requested budget and time frame to be agreed on by state, regional, and local authorities.

AI-enabled services show promise for gathering and analyzing stakeholder input through e-participation, web-based sentiment analysis and AI chat. AI-based tools show further capabilities in performance indicators' data collection, simulation, and analyses of the feedbacks generated to help planners and decision makers cope with the complexity and increasing uncertainty of urban environments. Simulations built with urban digital twin capabilities and high-performance computing, potentially embedding scenarios with uncertainties on model parameters and adapted to urban macro and micro-dynamics at play during ordinary days and peak situation, allow running structured sensitivity analyses, confrontation of alternative solutions, and considerably enhancing the interpretability of the insight generated. Moreover, the

potential integration with visualization and serious gaming mechanisms increases user involvement and understanding of the results.

23. Global Collaboration in AI Research

Internet was invented to help people communicate to efficiently and effectively perform tasks and solve problems. It has become the backbone of just about all global collaboration in research and development. Major AI developments rely on massive talent and resource sharing, mostly through this global information exchange network developed to facilitate resource sharing in science and engineering, but also increasingly outside science and engineering, to be used increasingly in media and entertainment. Open sharing of knowledge in both process and substance is a pillar largely accepted as effective and desirable in both practice and ethics. Funders have embraced this ethos, most recently with the shift to create and disseminate open source large language models.

We pioneered one of the earliest global collaborations embracing the open sharing ethos in AI, long before the internet: the original Generalized Problem Solver project. GPS was intended to uncover the common base of human reasoning - finding solutions to the grand task of "problem solving." The first GPS program was a major expansion of previous monographs describing a "General Solution of Two-Person Zero-Sum Game," an analysis of game-theoretic strategies for two-person games with no ties. Our initial implementation uncovered the hidden determinism of fictitious play best-reply strategies controlled by a single global optimizer, through rapid exploitation of the opponent's approximation. Time feedback from actual play improvement dictated the best response strategy to exploit the opponent. While GPS had a mighty impact on the global research landscape, importantly, the research community espoused the open sharing ethos before and carried it forward after.

24. Monitoring and Evaluation of AI Projects

Introduction to Cognitive AI for Societal Resilience is written primarily for practitioners working at the 'implementation' stage of using cognitive AI for disaster preparedness and response. This stage is determined by the decisions that practitioners take on the use of cognitive AI to make contributions to social outcomes. In particular, significant questions of 'monitoring and evaluation' arise. The contributions that cognitive AI can help make are part of a 'project-level logic' that spell out how using cognitive AI has relevance for operation. Importantly, social outcomes are a meaningfully defined interconnected set of characteristics of people's lives and societies in which they live and as such cannot be defined purely according to procedural definitions of specified things counted as 'output' measures. Rather, practice needs to distinguish between outputs and outcome measures. The challenge of

using AI for societal resilience is then to see how at the project level, interventions using cognitive AI will help achieve the objectives of the specific project. Linked closely to this question are judgments that practitioners make on fundamentals of the AI approach to be used, such as the degree of AI autonomy in the project system, the types and sources of data used, and the data representational techniques employed. It is noted that, generally, resilience appears to be understood in special connection with reference to an understanding of 'physical risk'. However, the insights in this chapter and its focus on the data inputs, functionality, and learning, applies equally to the establishment of cognitive AI projects for other kinds of activities that go into disaster preparedness and response.

25. AI and Environmental Sustainability

The environmental impact of AI technologies is both a concrete and pressing problem, at risk of worsening in the next decade. Typically for general purpose technologies that are diffusing fast and large scale, the negative benefits generally manifest as energy use, carbon emissions and e-waste on very large scales. From self-driving cars to biometric surveillance, AI systems entail enormous processing costs and usage. Moreover, the power demands are not only for inference but also training the models which can require hundreds of Terawatt hours, roughly equivalent to the power used in the entire state of California in a year, for a single major natural language processing application.

Yet if harnessed and deployed well to particular ends, AI can also help mitigate three major forces working against sustainability: (1) the sheer scale of human activity; (2) the fact that a large share of that activity is wasteful or otherwise dangerous; and (3) the uncertainty that cloud solutions give us about the technological course the economy will take. Some AI experts have highlighted fields that can potentially use AI for sustainability. One of these is energy distribution, which already depends on forecasting demand and supply from renewable sources. Using AI for demand side management decisions can also make a difference for balancing electric grids. AI can optimize building energy consumption and dispatching residential storage batteries. According to estimates, managing up to one-third of battery and electric vehicle charging loads would obviate the need to construct truly enormous amounts of excess capacity of the electricity grid. Other estimates suggest that the UK could gain significant avoided costs in 2030 by using AI to optimize the charging of electric vehicle batteries.

26. Conclusion

The proposition to consider for the development of cognitive AI is that cognitive AI must be developed as AI embodied into systems whose design, architecture,

capabilities and intended functions arise out of a detailed view of the physical-external and societal-internal, constraints-and-opportunities assumptions of these systems and of the interactions between these two complementary perspectives. We have shown the central role that cognitive AI should and can play in the development of resilient systems and have identified key features of cognitive AI that render it uniquely suited to this development. We have considered these features in detail, paying special attention to the question of how cognitive systems – and by implication cognitive AI-enabled systems − should be built. Addressing this issue of system-building has led to identify an important class of cognitive functions pertinent to designing cognitive AI-enabled systems for societal resilience. These are anticipatory, communicative, interactive, trustworthy, and open-to-learning functions. We have pointed out that what we consider the distinguishing feature of cognitive AI-enabled system for societal resilience is the inclusion in the design and architecture of the system's heterogeneous AI components of deliberative processes that make use of Anticipatory Cognitive Maps implemented as a hybrid cognitive hierarchical architecture. A third distinctive character, that cognitive AI-enabled systems for societal resilience must have, derives from these systems being meant to carefully intertwine AI capabilities with human capabilities to pursue jointly – as closely interdependent engaged partners – goals achieving systemic resilience.

References

[1] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.

[2] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.

[3] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.

[4] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.

[5] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.

[6] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.

[7] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).

[8] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.

[9] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[10] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[11] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[12] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[13] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[14] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[15] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[16] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.
[17] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.
[18] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.
[19] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.
[20] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.
[21] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.
[22] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.
[23] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[24] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.
[25] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.
[26] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing

healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[27] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[28] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[29] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[30] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[31] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.
[32] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.
[33] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.
[34] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.
[35] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.
[36] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.
[37] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.
[38] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.
[39] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 2: Theoretical Foundations of Cognitive and Neuro-Symbolic Artificial Intelligence

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Cognitive AI

Cognitive AI is an evolving, metastable, multidisciplinary research field. Its goal is to design and implement artificial systems endowed with intelligent cognitive functions, and the power of causal understanding behind them. Instruments made of computational and mathematical models of natural cognition and neurology, and advanced algorithmic techniques, are applied to solve the open theoretical problems in this field and develop applications solving nonlinear, ill-posed, high-dimensional, and computationally intractable inverse learning and decision-making problems [1-2]. Essential Cognitive AI principles and tools are causal modeling and reasoning by relations; symbolic representation, processing, and reasoning; relations-based generalization and analogy making; neuro-symbolic inference automation; multimodal object learning; and relational deep learning. Entries to the evolving Cognitive Systems' core are the theory of causal learning and causal modeling from general, combinatorial, statistical, and logical perspectives; the principles of object, scene, and event semantics; their hierarchical multimodal relational structure; and deep learning as representation learning according to this hierarchical structure. Specialized domains of the developed Cognitive AI technologies are multimodal perception; 3D object and scene understanding from photos, videos, and robots' proprioception and exteroception data; grounded syntax and grounding in creativity from random texture generation to visual question answering; scene and event understanding; situated AI; natural language understanding and benchmarking; neuro-symbolic deep learning; neuro-symbolic planning; and testing and verification [2-4]. Special tools for Cognitive AI are relational neural networks, causal graphical models, cognitive and neuro-symbolic architectures, and cognitive visual tools. Relationships between Cognitive AI and

Cognitive Modeling, cognitive robotics, Neuromorphic, Neurolinguistic, and Symbolic AI, Statistical and Causal Learning, Infocommunications, and Trustworthy AIs, and the foundations of Intelligent Optimization are discussed.

2. Deep Learning

Deep Learning is one of the most prominent AI technologies and is shown to achieve state-of-the-art performance in many AI applications. Deep Learning is currently implemented as Deep Neural Networks that are function approximators, typically with additional hierarchical layer structure inspired from human-brain processing [5-6]. The earliest Deep Learning algorithm applied to AI is a restricted Boltzmann machine. Various Generative Adversarial Networks (GAN) variations introduced later also achieve generative performance in Computer Vision or speech and audio synthesis tasks. However, the vast application of Deep Learning as function approximators in AI is under Feedforward-NN, Convolutional-NNs, and Recurrent-NNs that solve a great many classification, detection, recognition, or translation tasks using supervised learning with large labeled datasets.

Deep Learning techniques have enabled the realization of Intelligent Vision Systems and Intelligent Speech and Language Systems, which were considered as enormous technological frontiers to overcome [7,8]. These Deep Learning models are now as much like software engines, driven by the massive amounts of relatively clean and labeled data amassed on the Internet, as by the artful constructors of the engines. In addition, AI applications were simplified or built based on specialized learning pipelines for different tasks, not designed as a general-purpose architecture: Modules, Task graph, and Learning pipeline.

Deep Learning is so powerful because it exploits three major Leverages to enable the use of Neural Networks with much larger model capacity. The first is richly-structured Neural Networks with many more layers and with easy access for many more connections [9-12]. The second is Data, Data, and Data: large-scale labeled datasets are readily available, such as ImageNet, Text, and simultaneous transcription and translation corpus. The third is computation. Algorithms, like the stochastic gradient descent with back-propagation, allow the use of larger models, trained on larger datasets, using graphical processing units and cloud computing.

2.1. Fundamentals of Deep Learning

Deep learning is a data-driven, problem-solving approach based on engineering neural networks and unsupervised learning functions, generalizing the founding principles of neural computing and self-organization [7,13-15]. It is further motivated by brain theories, and so increasingly incorporates biological design features. Cognitive

Problems involve solving challenging transformations on symbols, such as transducing a sentence in natural language into a logical formula representing its meaning, transducing a formula representing a worldly situation into a natural language sentence, or optimizing a symbolic logical formula, a graph, or a set thereof. Deep Learning architectures, algorithms, and applications for Cognitive AI, Neuro-symbolic AI, Neural-Symbolic AI, are better understood involving the overlapping subfields of Algorithmic AI, Cognitive Neuroscience of Language, Neuro-Linguistic Programming, Neural-Symbolic Cognition Research, Symbolic AI, and Statistical AI.

In the short period of twenty years that have passed since the publication of seminal Neural Computation papers investigating Learning, and Processing, structured representations using Artificial Neural Networks, deep learning and its applications have surged into the forefront of AI. Human-level performance has been achieved in large-scale image classification, visual object detection, and speech recognition problems [9,16-18]. Similar levels of performance have been achieved in some areas of Natural Language Processing, such as Language Modeling and Machine Translation. A parade of industrial applications have been deployed, from visual face recognition to automatic tax analysis, or financial fraud detection, to mention a few. Deep learning is also being integrated into other Cognitive AI disciplines, such as Computer Vision and Knowledge Representation, and Reasoning. It is being used to retrieve images of a desired type, or from a given geographical area, or to perform detection of high level symbolic entities in vision, such as, for instance, institutions in rural landscapes.

2.2. Applications of Deep Learning in AI

Deep learning is a tool for automatically discovering patterns from data. It is currently being applied in AI, revolutionizing many tasks across many fields, achieving impressive new results. DL applications in general intend to create an intelligent agent that mimics some behavior associated with humans or machines. For example, applying DL for solving behavior-based tasks such that the behavior is associated with performing an intelligent function in a real world scenario [2,19-20]. DL is also applied in natural computation processes such that it manages cognitive tasks related with cognition in embodied minds. For example, the process of reasoning, simulating, envying, i.e. mental state attribution, etc. Socially instructed robotics tasks are also being solved with DL in areas of human Robotics interaction and HRI. Communication and emotions are also used in DL for behavior-based tasks such as chatbots or conversational models. DL is also being applied in both social cognition and action recognition tasks in social robotics.

Applications of DL in general are aimed to create neural networks or models that solve behavior-based intelligent functions for non-cognitive, cognition and socially

instructed robotics tasks. Considering the above premise, DL is being applied to different types of intelligent functions in AI. A significant going on area is the application of DL for computer vision [9,21-23]. Thanks to GPU computing, large scale data, and multi layers neural networks, models have been created that are able to associate and classify images with an error for a well known data set that is lower than the human standard error. Models were also applied to solving the object detection problem.

2.3. Challenges in Deep Learning

There are several major problems with current Deep Learning: (1) It requires huge amounts of labeled data for all tasks we want to solve, which is a problem for many applications. (2) It requires inefficient and long end-to-end learning of specific tasks, though progress on meta-learning, self-supervised learning and transfer learning will help. (3) The learned models, which are neural networks, are very inefficient for both learning and for end use, and it is unclear how to merge knowledge-based or logical methods with Deep Learning and (4) As evident from being fooled by adversarial examples, these models cannot have explainable, interpretable and reliable behavior [24-26].

Current deep learning models depend on data-driven learning from large amounts of annotated real-world examples, with the user providing only the labels for a few examples of the input output mapping for the particular task. There are two major limitations: (1) any real-world AI system should be able to learn from a few examples, or even from no labeled examples and instead leverage prior knowledge and similarity among categories, and (2) for many application domains, there may be no large collection of annotated real-world corpora available for training. For narrow tasks, such as speech recognition in a domain with a large collection of transcribed signal data available, deep learning works well [8,27-30]. But for many other tasks, labeled data is not available. For example, to build AI systems capable of multi-modal dialogue or visual video understanding, with the capability to learn about anything that we tell them, we cannot collect such massively-scale corpus data, since those tasks are too complicated and open-ended and, more importantly, it is prohibitively expensive to collect such labeled examples from humans who may also need to spend hours annotating video before coming up with a small number of labeled examples.

3. Reinforcement Learning

Automated decision making involves choosing the best option for a beginner or learner through a series of events or trials. This is the basis of Reinforcement Learning (RL), which uses positive and negative reinforcement to occupy decision-making gaps in the realm of AI. Using the principles of behavioral psychology, RL has its own taxonomy

to evaluate across multiple dimensions, both in terms of usability and performance. Its building blocks are also used in Neuro-Symbolic AI Building Blocks: Learning How? Whenever we start to interact with the environment, in our initial interactions there exists a notion of browse-observe-execute. We execute an exploratory action that allows us to observe the state of the world [9,31-33].

RL uses the "reward" function for sense and direction. Rewards also allow us to understand the proximity of any action to a desired plan. Over time, with the experience of previous interactions, we can reach the problem of resource optimization. The function allows us to predict the future returns of actions - thus showing us the exploratory path to follow [34-36]. RL promises to tackle problems of high complexity, where the search space for the best plan is huge and basically invariant with respect to the parameters of an optimization method. The high complexity of the search space does not allow for an exhaustive exploration of the possible paths that lead to a minimum of loss/cost, associated with a measure of proximity to current use. Nevertheless, there are many problems for which reward calibration - or indirect observation using lower-order functions on the entire knowledge - is possible. Furthermore, in view of the principles of first and second systems of simulators, it has been possible to develop and introduce approximation functions that delicately balance prediction capabilities and the number of free parameters.

3.1. Basics of Reinforcement Learning

Reinforcement Learning has become a central research field within Artificial Intelligence and a cornerstone of current advancements. RL is a computational approach to the fact that the ability to find and exploit causal relations is very instrumental for intelligence, and in many tasks it suffices to learn by making and verifying guesses about what actions will produce favorable consequences. RL has evolved from its origins in Behavioral Psychology via the principles of Conditioning and Operant Conditioning to become a mathematically rigorous framework, with its own foundational ideas, that serves as an umbrella for models and algorithms from various institutions worldwide [3,37-39]. A powerful feature of RL is that it can be used to find solutions to complex sequential decision-making tasks for which written algorithms are either unknown or too complex to be feasibly defined by a human.

Within the RL framework, an agent employs a feedback signal, the reward, to find solutions to complex tasks by trial-and-error, hence learning from the consequences of its actions, irrespective of the goals, or utility function, of the agent designer. Moreover, reward functions are often easier to define than utility functions; an Artificial Intelligence designed to autonomously gather information would cheat,

rather than act to get as much information as possible, if it had a utility function built on knowledge! The learning task is modeled as a Markov Decision Problem, which contains a probabilistic state transition function and a transition reward function. The model choice is justifiable when the design of the reward function is largely decoupled from that of the state transition function, and it is appropriate for a wide range of tasks. Interestingly, it is also known to approximate many classes of temporal-difference learning tasks as well as inverse reinforcement learning tasks.

3.2. Key Algorithms in Reinforcement Learning

Among the various ways to frame an RL model, the Markovian Decision Process (MDP) formalism is the most prevalent in the literature. An MDP is defined by the tuple (S, A, R, T, γ), where S is the state space (or, more accurately, the state space of the environment); A is the action space; R is the reward function; T is a transition function defining a probability distribution over successor states for each state/action pair; and γ is a discount factor. The environment is subject to Markovian assumption, that is, the current state dictates the distribution of future states, and is memoryless. The agent receives a state s, selects an action a according to some mapping π, which results in an instantaneous reward r according to R(s, a) and sends the action to the environment that results in transitioning to a new state s′ according to T(s, a, s′). This loop continues to execute until the termination, when the agent receives a special (non-final) state that disables any further action.

In short, RL solves the problem of learning the mapping π that maximizes the expected return E[$\sum_{t=0}^{\infty} \gamma^t r_t$]. During learning, RL algorithms must balance exploration (testing unvisited states) and exploitation (taking actions that maximize the return). In its most common variant, RL refers to situations where the transition function is unknown (it is a stochastic process) and tries to learn both the policy π and the transition function during the interaction loop. From the perspective of exploration, RL encompasses both the cases of active and passive learning, with increasing levels of exploration as the agent collects samples from the environment. From the perspective of the used model, RL algorithms may be characterized as model-free, model-based and Dyna-style.

3.3. Real-World Applications of Reinforcement Learning.

Reinforcement learning has recently made inroads into a variety of real-world problems. Its most famous successes have been two breakthroughs. One was the success of deep reinforcement learning on playing video games from pixels and generalizing across many of them [36,40-43]. The other was the success of a system which, with a combination of deep reinforcement learning, supervised learning, and tree search, outperformed the best human players. Other RL successes include game

walkthrough automation, in which deep learning from a large dataset of games is followed by RL to achieve human-level performance; as well as recent breakthroughs on the complex multiagent environment for AI and on the long-term sequential decision-making problem. A core algorithm was behind these advancements. It utilizes the classic Monte Carlo Tree Search algorithm in conjunction with deep reinforcement learning.

Driven in part by this success, RL's model-free techniques are being adapted to many fields outside of traditional sequential decision-making. Recent successes include autonomous racing, robotic motion control, architecture search via reinforcement learning, hyperparameter optimization via reinforcement learning, intricate sequence generation problems like text and protein generation in the context of deep learning, recommendation systems, algorithmic video editing, and algorithmic financial trading. Other simpler applications are credit assignment problems, like incentive design. This success can be attributed to factorized structure that can leverage exploration and the nature of the training problems.

4. Causal Reasoning

Causal reasoning allows humans to interpret and model events as based or relying on other events, linking them together through Cause-Effect relations. A Cause-Effect relation provides a way to link things happening in parallel and/or at different times; allows us to manipulate things in order to influence others; provides an understanding of things around us in terms of explaining why things are the way they are; and is the basis of language semantics providing the background for the truth and correctness of the sentences we speak. It in part allows us to reverse engineer awareness – and considering its important function in our cognitive and symbolic intelligence, it is important to model it in AI systems.

Academic research has presented various models of causal reasoning in AI. The most discussed one is the model of counterfactuals, which explains that whenever we can imagine how changing the value of variable X to a different value from the one it would normally take would also change the value of variable Y to a different value from the one it would normally take, and especially influence a Y-value belonging in the set of values selected out of the probability by the intervention effect in the Ask-Why sense, then we say that X is a cause of Y; and this allows us to model by the causal Model of Variables of the Causal Structure of the data-points of the problem domain (ordinary cases) and/or by the Intervention Graph how those probabilities would be modified for those particular cases through an external manipulation.

4.1. Understanding Causality in AI

Causal reasoning has received recently much attention in research fields of different disciplines and research areas, such as philosophy, statistics, economics, and AI. This is mainly because understanding causation, and thereby the systematic modeling of causal relations, is a necessary requirement not only for humans to survive, but also for an autonomous agent that behaves appropriately in a real-world environment. For a long time, researchers in AI followed the philosophy of simulationism: creating systems that simulate (parts of) the human mind in order to understand it. Not surprisingly, the main goals of many AI systems, especially the early "classical" systems, might be better described as "minds in a vat" — systems with a focus on pure cognition or perception — rather than intelligent agents that interact with the external environment. Hence, the agents' internal cognitive models of the world need not only describe the structural relations of the different components, but also their causal dependencies and hence, allow planning and acting.

More recently, the predominant trend of modeling and training AI agents on currently available data, has led to a more task-oriented paradigm shift. Nevertheless, the growing concerns regarding the still poorly understood abilities, the reasoning and explaining capabilities of current data-driven or deep learning-based systems, have shown a renewed interest in cognitive and neuro-symbolic approaches. The by far most useful task in this regard is the building of a causal model, either directly used by the AI system itself or as additional system knowledge that is exploited during the information processing tasks on the AI system. Following the counterfactual- or causal-modeling approach, the essential task of such a causal model is the representation of 'actually happened' and 'would-have-been-happened' relations to allow contrastive inferences — the modeling of structural relations of multiple events happening and observed at different times and places in human or agent life.

4.2. Models of Causal Reasoning

Causal reasoning has been an indispensable part of human intelligence, as humankind's understanding of the world has evolved from the basic empirical notion of correlation to the keen ability to act upon and thus generate motion in nature. Contemporary Artificial Intelligence applications, when not physically embedded in the real world, rely on Machine Learning models that analyze data and models of abstract representations that encode different aspects of the world. However, neither of the proposed approaches is capable of achieving the complexity and variety of forms of human intelligence. The semantic turn in AI, in general, and the emergence of Cognitive AI, in particular, steer AI back to looking for the sources of skills and abilities that are more symbolic in nature and enable intelligence that is not limited by

a given pre-programmed set of tasks. Yet, without the capacity for dealing with open-endedness, the evolution of intelligence - artificial or not - must inevitably stall. Theoretical models of open-ended growth in complexity, and the interplay between memory and working memory in enabling this growth, are paving the way for Artificial General Intelligence.

In the real world, there is a differentiation of actions that activate certain parts of memory while bypassing others. Memory-guided reasoning over a pre-flow of events enables action anticipation, revealing the nested structure of action sequences and their functional role in motion generation. A special case of nested recursion is a system of ordinary differential equations - it describes a whole class of motion sequences. This theoretical window reveals why certain temporally-augmented decision-making/action selection and synchronization problems become easy when modeled as dynamical systems.

4.3. Applications of Causal Reasoning

Here we provide a selection of recent demonstrations of causality in research that could be considered the most important applications of causal reasoning at the moment. Applications of causal reasoning were traditionally concentrated in the areas of science, natural and social. Research in these areas typically seeks to explain data generated by natural processes – a group of individuals of a species living in the same ecosystem for a certain time, and data generated by the interaction of particular social groups – families living during some time in a particular area, their demographic trajectory, their marriage and reproductive behavior, their economic exchanges, etc.

That said, when people hear about applications of causality in AI, they mostly think about causal representations of the world and causal reasoning in robots and autonomous agents. Such robots would acquire structured causal knowledge from the environment, as human toddlers seem to do, and progressively refine these structures by acquiring finer details of the relations that their interactions with the world show us. The acquired knowledge structures would allow these robots to carry out causal reasoning, including causal and explanatory reasoning in perception, conception and conception planning, as well as action, tasks, and manipulation tasks. These robots would have practical causal knowledge about the physical world, which they would use for effective, efficient, and skillful action in their continually ongoing interactions with the environment.

5. Neuro-Symbolic Integration

A variety of cognitive capabilities are still beyond the ability of current AI techniques, particularly large deep learning models. A notable example of this is syntax in

language, or analogical reasoning, which is both an expert domain of current human cognition as well as a fragile point of our current AI systems. Cognitive Science theories motivate integration between symbolic systems and sub-symbolic neural architectures to leverage their joint strengths while compensating for weaknesses.

The relationship between symbol-based and sub-symbolic representations has been debated for decades. Neuro-symbolic AI leverages the best of both worlds to produce capable systems. Neuro-symbolic integration, or hybrid architectures, combine the strengths and weaknesses of symbolic and subsymbolic AI systems, yielding more comprehensive models of cognition and potentially more competent AI systems. Integration possibilities can range from simple interaction between two independent systems to a single architecture that uses both representations in a complementary way.

Neuro-symbolic AI promises a seamless integration of neural and symbolic components, resulting in improved capabilities in areas such as reasoning, explainability, generalization, and learning efficiency. Such integrations may also produce cognitive systems with attributes like compositionality, and language and knowledge grounding that are typically elusive in pure neural systems. Various forms of neuro-symbolic AI occur at different levels of abstraction in different modalities. The understanding of human neuro-symbolic abilities sheds light on the right balance humans have achieved via evolution. While pure neuro-symbolic designs may not be optimal, tweaking symbolic knowledge back and forth to sub-symbolic systems can increase overall efficiency in task planning, learning, and execution.

5.1. Overview of Neuro-Symbolic AI

Neuro-symbolic AI is a recent malleable umbrella term that builds a research space that connects various neuro-symbolic work, where neural and symbolic technologies are exploited to develop phenomena that would require both types of technologies when developed in pure form to be explored with the objective of joint capabilities. The degrees and types of possible integration between neural and symbolic AI are of many forms [5,6]. These connections include the integration of the modules of distinct or not models based solely in neural or symbolic processing, the use of symbolic representations inside neural models, as well as the incorporation of machine learning to the learning or inference process from symbolic representations. Despite the diverse types of possible integration, they aim to explore the joint expressiveness and capabilities of neural and symbolic AI.

Neuro-Symbolic AI is a multidisciplinary challenge that exploits resources from different areas of AI, where cognitive neuroscience may be used to inspire aspects of its formulation and different philosophical theories may provide support to the claims of its proponents. Besides developing models that operate under the theoretical pillars

of Cognition and AI areas, this multidisciplinary challenge also raises questions and provides foundations for theories that connect the themes with the ultimate goal of understanding and modeling computational aspects of the human mind. The proposed NS-AI approaches use as neural components some form of deep learning like, for example, deep learning used in vision or language processing tasks.

5.2. Benefits of Neuro-Symbolic Integration

There are at least three main benefits of neuro-symbolic integration. First, with regards to compositional generalization, neuro-symbolic integration may help alleviate the notorious issue of compositional generalization in neural-symbolic systems. Second, neuro-symbolic integration may help increase the robustness of neuro-symbolic systems. In particular, it has been suggested that systems that do not only learn from or make use of the statistical regularities present in the input data but also use symbolic world knowledge about the relations present in the environment that are not necessarily present in the input data may perform better in transfer learning. This is particularly true for cases in which the data have statistical regularities that are too weak to allow the neural components of the systems to learn them effectively. In this regard, we remark that symbol-level reasoning is important because it goes beyond the simple computation of pattern correlations induced by statistical learning. Overall, the usage of neuro-symbolic systems in real-world applications that are affected by safety and security issues like self-driving cars has been repeatedly encouraged.

Last, about the second benefit, neuro-symbolic integration may be vital for the realization of AI systems that yield physically correct, grounded or factual, and verifiable answers to human users, similar to what traditional symbolic systems do. Indeed, this element has been at the basis of the original pursuit of hybrid systems as it is also key for achieving a high level of explainability of the systems' predictions, given that humans often want to have an understanding of how a machine made a certain prediction, given a certain piece of data. In this regard, we note recent crucial advances the field has made in the definition and development of explainable neuro-symbolic systems.

5.3. Structured and Explainable AI

The question of structured AI, also known as explainable AI, refers to the obvious possibility to combine neural AI with symbolic reasoning, since we know that neural algorithms for perception can be pointed at specific symbolic tasks. However, we also know that perception alone is not sufficient to achieve anything that we consider general intelligence. Even in the case of pattern recognition and decision-making, arguably the two areas where NNs are today at their best performance, there is still much about their operation that we would like to understand and make more

predictable. Probabilistic graphical models such as Bayesian networks, CRFs, or Markov Random Fields and Decision Trees are very popular, but they are bottleneck models that use local approximations to represent the whole structure. A typical tree may represent a pattern with about 30-40 variables while structured models can treat several million features. Symbolic does not mean deterministic but deterministic structures can be enforced on neural networks, and also probabilistic structures at the output level. A number of hybrid approaches enforcing such abilities in different ways have already been developed, for instance symbolic sectors and STNN architectures.

The question of interpretability is by no means new and many statistical and machine-learning linguists have faced it since the mid-nineteenth century, when the confusion between rationality and reality became mathematically clearly pointed out: it is not the method that justifies the theory, it is the theory that justifies the method. The typical way to find ideas for methods is to look at nature.

6. Decision-Theoretic Frameworks

6.1. Introduction to Decision Theory

Decision theory serves as the foundational framework for understanding reasoning and decision making in countless models of intelligence, whether human, animal, or artificial. Broadly speaking, decision theory encompasses any formal modeling system that tackles the problem of figuring out which action is best to take. When someone uses the phrase "a decision-theoretic model", they could be referring to any possible theory covering this certainly enormous catalog of functions. When we use this phrase, we are specifically referring to theories that take prior action selection probabilities as inputs and then output the probability assigned to being in each possible world, after normalization and possible approximation, and which are based on a utility function for the exponential approximation of that normalization. Decision-theoretic models are the most common models of decision-making in cognitive science and in AI, and are often designed to be directly implemented in cognitive or neural networks. These models are called decision-theoretic because they rely on decision theory as their core modeling premise and its output as input. We will also be limiting our focus to internal decision theory; that is the theory of apparently rational behavior that does not involve physical action. A practical definition of "theory" is that it specifies a computational function. The first input-output map that we would like decision theory to implement is that of being in a world. More formally, the task is to approximate the

6.2. Key Decision-Theoretic Models

Are optimal predictions of the world while being conditioned on action selection probabilities or a related quantity. We also want the model to be able to deal with all

possible variables. The primary function of decision-theoretic models beyond action selection is making predictions about how an agent will behave under its decision theory. A useful model for this purpose can be created by assigning a direct probability to each possible assignment to the random variables of interest. In order to normalize properly, we can decompose the joint distribution into the product of the conditional distribution acting as the normalizer for the final expression times a prior over action selection probabilities that is factorized over individual action probabilities, though that factorized form need not necessarily be the prior over action selection probabilities for correct actions.

6.1. Introduction to Decision Theory

The essence of autonomous action comes down to making correct decisions given the available knowledge. And the field of decision theory attempts to provide a principled foundation for our ability of rational behavior. Decision Theory is a unified formalism for modeling rational behavior regardless of the details of the reasoning or inference involved. It is concerned with the conditions under which an intelligent agent's choices can be said to be rational, and whether it would be reasonable to assume that an intelligent agent intends to maximize a utility function over its choices. Broadly, Decision Theory provides a way of modeling the relation between the intentions of an intelligent agent and the options available to it, by suggesting that rational agents always choose the action that maximizes the expected utility of the system, based on a prior model of how the actions will affect the expected outcome. Expectations and intentions are indeed crucial for an agent's decision. We assume that agents strive to attain goals. They evaluate their expectations of attaining goals and make their choices according to these evaluations. The simplest model of decision theory identifies goals with numerical utility functions. It assigns a single scalar value to the overall desirability of being in a certain state, or of making the world change in such a way that the agent is in that state, after the action has been applied. A typical use of decision theory in AI is to model an agent's behavior given the action probabilities it produces. In this view, action probabilities are influenced by the agent's degrees of belief and degrees of value – its predictive and normative aspects, respectively. Decision theory plays two vital roles in AI. It attempts to formalize the meaning of rational action, and it explains how rationality influences action.

6.2. Key Decision-Theoretic Models

Decision theory has two distinct but related branches. Bayesian decision theory focuses on agents making decisions based on uncertain internal and external world models. Game theory considers interactions among multiple agents, with each agent making decisions that affect the likelihood of states of the world model of all of the agents. The

social agents then pursue their policies to maximize their respective utility functions. The two branches are strongly connected. Building on its concept of a common prior probability distribution over hidden variables, Bayesian decision theory can be extended to a multi-agent case to analyze the motivations and decision making of each agent. The decision problem for each agent is to maximize its expected utility based on the current belief state of all the agents from the prior joint probability distribution. In game theory, the prior probability distribution is a special case. Agents are attempting to maximize their payoffs based on maximizing or minimizing the payoffs for other agents. Thus, game theory assumes the agents have common knowledge of the parameterization of the payoffs.

An additional, intermediate-level modification is the use of correlated equilibrium (CE) concepts. CE allows utilities to be discussed while also considering the need for a specific joint probability distribution. In particular, CEs connect Bayesian and non-Bayesian cooperative decision making. These three models serve different purposes in AI systems. Neuro-symbolic systems with Bayesian modules can respond in very specific situations. Non-Bayesian game-theoretic systems can create large models or ensembles that approximately match utility functions, even with early, hidden incorrect agent assumptions. In contrast, systems with CE concepts can do both relatively efficiently. In particular, CEs allow the prediction and explanation of human behavior conditioned on specific internal models for social interactions.

6.3. Applications in AI Systems

One broad area of application for decision-theoretic models is in automated planning. Models that treat the difficult portions of the planning task as computation rather than representation have only recently been argued to be superior. The most visible of these models is the factored Markov decision process that clusters together similar patterns of behavior rather than merging states in a Markov decision process. This technique is borrowed in a simple version from connectionism, where the patterns of activation are abstracted, but the similarity metric used is not trained. Much of the recent success of reinforcement learning algorithms on complex real-world problems is now due to connectionist function approximation. The attractively simple reinforcement learning model can be generalized with respect to richer parametric classes. This generalization appears in other recent work in planning using nonlinear programming to minimize expected time under continuous-time Markov decision process models.

The work in the previous paragraph describes quite essentially model-learning models. These are duals of the model-use systems. Model-learn systems need clear instructions on what to observe but they can operate under spotty, noisy sensors where model-use systems need sensors with good accuracy and coverage, already in the planning model

in order to function well. These AI systems do not need to be goal-directed in the sense that they must go for gold. In fact, these operations would set model-use planners at loose ends but these model-learn systems can be autonomous observers. No models of the goal are needed for these systems to operate. Model-learner systems with a planner operational with simple parametrically small processes continue to be rarities.

7. Human-in-the-Loop Systems

The key aspect of AI that differentiates it from traditional IF-THEN symbolic architectures is that AI learns by itself from user interactions. In general, the space of Natural Language Symbolic programs such as regression, classification, prediction, etc. is too large to be covered using prior knowledge alone. The lack of an Axiomatic Prior or the impossibility of covering the infinite space of symbolic programs given the finite set of training examples that will ever be available are two arguments that show that human input in some form or another is essential for Intelligent AI to exist. Different paradigms such as reinforcement learning, imitation learning, cognitive behavioral techniques, etc. seek to cover that void by proposing a method for the user to express their preferences and how these preferences can be used to fine-tune the AI model. In the Cognitive and Neuro-Symbolic paradigm, the user inputs their background knowledge in the form of Neural Gateways — structures that translate specific programs from the Cosmos of Languages to a Neural Architecture that can be trained.

As Natural Language programs are complex and large-structured, the efficiency of learning can be greatly increased if the programmer uses knowledge that allows fewer learning iterations in a more optimized training process. The research community has therefore started to study all forms of User-AI interaction and how humans can help and guide AI Learning are some of its main goals. Therefore, different forms of Interaction Loop continuously analyze and infer user preferences and update in real-time the Neural Structure to guide and help the user with their cognitive load. Although still in its infancy, improvements in areas such as Neural Network Compression allow the creation of very light fine-tunable Lim Boot Neural Programs that can run efficiently on consumer and mobile devices. The creation of HIL systems has allowed a multitude of applications to flourish such as Smart Dialog, Visual Reasoning, Natural Language Generation, Code Completion, Knowledge Discovery, Test Generation, Content Creation, etc.

7.1. Importance of Human Interaction

The role of humans in the loop can be to supervise or cooperatively interact with automated agents. One of the most known applications of human supervision of AI are the so-called “data-labeling” agencies, where humans annotate datasets, which may be

used to train AI systems. The problem with this approach is that it is very expensive and tiring for humans, who usually do it without having a good understanding of what the AI is creating these datasets for. In fact, the end-to-end optimization of AI systems through the use of complex, risk-sensitive reward functions can necessitate huge amounts of human effort without guaranteeing useful results. Industrial use-cases of AI are also based on the idea that basic tasks of intelligent behavior can be outsourced to an AI agent for augmentation of humans. Many companies are developing AI tools, but many of the developed AI systems are still far from being capable of being completely autonomous, and they rather augment workers and empower them instead of replacing them. AI can handle large-scale analysis based on previously learned models and comprehensive datasets, humans can add context and understanding at higher levels of abstraction and take risk-sensitive decisions. Despite being designed to work independently, practical and real-life applications of AI usually emphasize human-AI collaboration, combining the inference ability of AI with the higher-level decision-making of humans.

7.2. Designing Human-in-the-Loop Systems

Human-in-the-loop systems function as tasks in which a machine performs most of the work while a human provides support. We theorize that a designer of such a task will be interested in answering two questions. The first question relates to performance: What is the maximum speedup or minimum time of completion for the task if we design it as a human-in-the-loop task? The second question relates to acceptance: Is the task acceptable to a worker? From a theoretical standpoint, these two questions are quite distinct and may even be treated independently. The speedup question links the two sources of performance to either task alone via the speedup factor of the parallel systems model. The acceptance question, on the other hand, is concerned almost exclusively with the human task. Intuitively speaking, we could break up the human task into a sequence of elements and design the human-in-the-loop task through an iterative process, testing each new element for acceptance.

Though these two questions can be answered independently and represent two important components in the design of a human-in-the-loop system, the relationship between task acceptance and performance for a worker functionalities needs to be well understood since bugs and shortcomings present in many proposed tasks can be traced to a lack of understanding of this relationship. The traditional assumption is that a human performs optimally on the whole problem that is acceptable. If the human performance is poor, the problem may not be acceptable. If the problem is acceptable, the entire problem is trivial, then it slows down the total performance. These comments should apply equally well to tasks wherein the machine and human work concurrently on the task and tasks wherein the two agents are working serially.

7.3. Case Studies of Human-in-the-Loop Applications

Recent advances in machine learning, and especially deep learning, have equipped engineers with tools capable of creating marvelous technological solutions. Machines and algorithms are beginning to outperform humans in highly cognitively demanding tasks, including the play of games such as Chess or Go. A neural network can outperform radiologists in the identification of diseased areas in X-ray images. Other results indicate that deep networks are capable of recognizing subtle signs of cancer in histological images of tissue, or quickly detecting markers associated with Alzheimer's disease in brain scans.

But for many applications, we still need humans to work with the machines, either to supervise them or to allow for reciprocal learning. Recent advances in linguistics, vision, and healthcare show that we can create increasingly intelligent services. Many intelligent systems require some degree of human input: to annotate training data; evaluate the effectiveness of recognition and understanding algorithms; correct application outcomes; discover new knowledge; or teach algorithms to recognize new concepts. Enterprises are investing in these intelligent systems. For example, users can annotate the political category of posts, which are then used as training examples for machine classifiers working to detect misinformation. A private sector firm has partnered to create a taxonomy of events that social media observers can annotate in machine learning systems. The intent is to build reliable and sophisticated algorithms for filtering spam using humans as a back-up mechanism.

8. Challenges and Future Directions

Abstract. Despite significant advances in cognitive and neuro-symbolic AI, for current models, there still remains room for improvement and important challenges to tackle. Here we map out current limitations and future research directions. In particular, we will focus on three future directions for research. First, we argue that more work is needed in the area of learning object and visual latent states in deep neural networks, as well as for modeling arbitrary differences among objects, visual congruence, and perform object-invariant recognition for understanding the visual concepts AI and vision systems in general rely on. Second, we argue that current cognitive and neuro-symbolic systems have serious limitations in performing complex differentially and temporally-constrained symbolic operations on symbolic representations. Third, we tackle the question of how to align symbolic representations and operations as performed by lower cognitive AI functions (like those of visual recognition, matching, and attention). Experiments indicate that current versions of deep neural networks are still far from actual objects in the visual domain. In fact, deep neural networks stumble notably on the basic ability of children and adults to extract and use visual object

representations with a set of rational properties. In particular, they struggle with the concepts of compositional structure, object individuation (the ability to identify and discriminate objects), distinct object identity and (canonical) object constancy over differences and perturbations in differing scenarios, spatial duality and relativity, as well as the fact that object recognition is a task that can (and often should) be invariant to certain kinds of variation both along the same and across differing instances.

The domain of cognitive and neuro-symbolic AI still has a long way to go before properly modeled general cognitive and visual operations can reflect models of true human intelligence. For starters, there remain questions about what "cognition" and "understanding" in the realm of different lower AI functions and general intelligence at the sensory to high cognitive function level should actually mean. However, it has become clear that the general area of not just symbolic AI, but neuro-symbolic AI in general is still lacking. In particular, what "cognition" and "aware" should actually imply in the realm of diverse lower cognitive as well as visual AI functions is still an open question.

8.1. Current Limitations in AI Approaches

At present, contemporary AI approaches are not able to achieve and reason about human-level cognitive complexity. Those results are not satisfactory for many application domains, like symbolic robotics, scene recognition and understanding, intelligent agents capable of autonomous problem-solving and decision-making in complex and uncertain environments, natural human-like language understanding and generation, social learning and interaction, interactive and multimodal learning, and many others. Above all, General AI is not at all achieved. The fundamental problem is that most of today's AI designs are overly specialized to the domains to which they are applied, with little shared structure among them. This specialization is not deliberate for the most part but is imposed by the fact that there is little or no good theory to guide the engineering of such systems and how to connect them together. Cognitive and Neuro-Symbolic AI can leverage cutting-edge advances in connectionist networks that are driving new results in image understanding, language processing, video-based activity recognition, dialog processing, and human action coordination.

Any learning system for intelligent behavior should include perceptual learning, language for communication, and analogical reasoning supported by symbolic comprehension, knowledge, and reasoning based on rules and relations. These three modules are fundamentally interdependent. Yet today's actors have enormous difficulty coordinating together. Sophisticated complex human behavior requires the integration of diverse modalities, including visual and auditory perceptual channels, character and object tracking, motion control, modeling environmental dynamics and causality, along

with verbal, gesture, and facial communications. No serious attempts so far have been made to integrate the leading focus area of learning and perceptual vision with cognitive reasoning based on symbolic understanding of language, causal, and relational knowledge.

8.2. Future Research Directions

Research on Cognitive and Neuro-Symbolic AI is at an early stage, which is exciting because there are so many potential avenues for future research. Our essay has concentrated on presenting a theoretical overview of existing symbolic and hybrid approaches. Theory can serve as a foundation for future research, especially benchmark results exploring capabilities vs. limitations of the set of currently available approaches to Neuro-Symbolic AI; Neurosymbolic approaches are also useful for benchmarking tests for human cognitive capabilities. These suggested future tests would help guide future AI research toward systems that can bridge the capabilities gap with humans, and would help researchers to explore areas already inside of human ability space, but generally lacking neuro-symbolic solutions or benchmarks. However, although our essay presents a theoretical foundation that is recommended for selecting the next training and benchmark tests, it does not explore endlessly deep subfields of currently available benchmarks and future potentials. Expansion into the space of potential Neuro-Symbolic architectural combinations of capabilities and research is a particularly motivation that should explore some initial promising directions. Identifying and exploring promising architecturally motivated Neuro-Symbolic combinations provides rich future potential both for building architecturally more intelligent systems and for enhancing understanding of the blueprint of human intelligence. These unique directions share initial research results with existing Neuro-Symbolic combinations, suggesting some possible collaborating paths. While the future AI directions are only a selection of a limited potential space, this cooperation and understanding offer the potential for the most interesting future work, whether inspired by human ideas or taking a totally different auxiliary path.

9. Conclusion

This work presented a theory of cognitive and neuro-symbolic AI that lays the foundation for the development of intentional AI systems. From Avicenna's conception of symbolic knowledge through Peirce's triadic theory of signs, Allen and Darden's mutual constraint conception of cognitive tools to Marr's functionalist theory of vision, we established the theoretical bases and principles that must constrict the implementation of neuro-symbolic systems that require the attribution of mental states to process and psychomimic behavior. The main plan of action was to defend the attribution of mental states to those systems from a metaphysical and epistemological

point of view. By proposing that intentionality can only be applied to systems that use symbol in Peirce's codependent sense and Allen and Darden's productive sense as cognitive tools, we provided a solid foundation for the attribution of practical intentions, conceptual intentions, perceptions, and beliefs to those systems. We also defended the attribution of cognitive mental states, explaining what they are by establishing the principles they are subject to and explaining how they are acquired. More importantly, the theories and principles proposed allowed us to develop a framework of cognitive intentionality in accordance with the most robust neuro-symbolic AI proposals available. With our theory, such proposals gain not only theoretical rigor but also explanatory power concerning the behavior of human and non-human cognitive agents. Finally, the proposed architecture can be implemented using currently available AI tools, something that we advocate for in future works. Summarizing, we developed an intentional framework for weak and inflexible AI systems, lowering, therefore, the ontological gap between AI products and human cognition. Finally, we note that no matter how much we work on minimizing that gap, we cannot forget what AI was designed to be: a simple tool meant to facilitate part of our day-to-day tasks.

References

[1] Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh TP. Neuro-symbolic artificial intelligence: a survey. Neural Computing and Applications. 2024 Jul;36(21):12809-44.

[2] Hitzler P, Sarker MK, editors. Neuro-symbolic artificial intelligence: The state of the art.

[3] Zhang X, Sheng VS. Neuro-symbolic AI: Explainability, challenges, and future trends. arXiv preprint arXiv:2411.04383. 2024 Nov 7.

[4] Thomas CK, Saad W. Neuro-symbolic artificial intelligence (AI) for intent based semantic communication. InGLOBECOM 2022-2022 IEEE Global Communications Conference 2022 Dec 4 (pp. 2698-2703). IEEE.

[5] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.

[6] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.

[7] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.

[8] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.

[9] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.

[10] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.
[11] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[12] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.
[13] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.
[14] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[15] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[16] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[17] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[18] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[19] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.
[20] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.
[21] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.
[22] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.
[23] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.
[24] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.
[25] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.

[26] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.
[27] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.
[28] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.
[29] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.
[30] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.
[31] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.
[32] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.
[33] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.
[34] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).
[35] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.
[36] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[37] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[38] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[39] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[40] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[41] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[42] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[43] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 3: Ethical, Legal, and Societal Considerations in Artificial Intelligence and Healthcare

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction

The emerging bond between artificial intelligence (AI) and healthcare is rapidly altering delivery processes in ways that promise to significantly enhance outcomes for patients and healthcare organizations alike. Healthcare practitioners and researchers are increasingly turning to AI for tasks that require intelligent solutions, with demonstrable success [1-2]. An historical example is a system that surpassed physicians in diagnosing cancers from analysis of medical records and images, and in performing analyses of some diseases. More recent examples include the deployment of an AI-based system optimizing the planning of radiation therapy for cancer, in ways that surpassed even the best human oncologists, and the creation of a software tool trained by deep learning techniques on a databank of retinal fundus images, that rivals human ophthalmologists in diagnosing retinopathy of prematurity from analysis of retinal images.

The increasing application of AI to healthcare brings with it heightened expectations concerning the contribution of AI to solving healthcare's most pressing challenges – achieving superior patient outcomes and increasing the efficiency of delivery while reducing unnecessary costs and the associated financial burdens on patients and governments [3-5]. However, grappling with these complex challenges requires the coordinated efforts of a range of stakeholders, including clinicians and clinical researchers, computer scientists, ethicists, and policy-makers, as well as patients and their advocacy organizations. An underlying premise is that AI solutions will be designed, validated, and implemented in truly collaborative fashion, with disciplinary and cultural perspectives from both healthcare and computer science being equally

valued [6-8]. Moreover, given the critical nature of the patient-clinician relationship in healthcare delivery, patients themselves must be engaged early and throughout the development process, to ensure that AI applications adhere to principles of health equity to not only optimize outcomes for affected demographic sub-groups, but also avoid artificial intelligence-induced harms.

2. Fairness in AI

In this chapter, we discuss the topic of fairness, which is perhaps the most explored ethical concern within AI. It is an area that has captured the attention of many researchers due to the increasingly important consequences technologies have on access to opportunities in sensitive areas such as hiring, credit assignment, law enforcement, offender recidivism evaluation, and healthcare [7,9-10]. Fairness is a broad concept with many definitions. Different definitions are required for different situations and to evaluate different forms of discrimination or inequality. Hence, measuring fairness is also a challenge. Finally, when people build and use AI systems, many challenges still remain to ensure fairness is achieved. Some of these challenges stem from the complexity of human decision-making [1,11-14]. Other challenges point to a lack of established practices in how to approach fairness: decisions made about definitions of fairness, how they are measured, and whether they are sufficient to achieve fairness in real-world consequences, are not always taken into consideration.

Because fairness is often defined as the absence of bias or discrimination against a certain group of people, certain existing ML fairness measures assess various aspects of bias. However, a recent study on fairness in healthcare ML studies suggests that authors do not engage with fairness issues consistently or sufficiently [13,15-17]. Some report only disparate treatment on various protected characteristics or just report model performance disparities conditional on protected characteristics. Others focus exclusively on model performance disparities and completely ignore disparate treatment. Others report disparate treatment biases but apply mitigation strategies regardless of any established fairness definitions. Hence, there seems to be a lack of consensus on how fairness, bias, and mitigation should be defined and measured, which can contribute to more or less fairness issues in the use of healthcare ML technologies.

2.1. Defining Fairness

An increasingly common refrain in conversations about AI and other algorithmic decision tools in healthcare is that they need to be fair or must be designed to minimize the potential for bias. While the appeal to fairness is almost universal, it is not always clear how it is being defined or what intervention is being called for. For example, the claim that an algorithm is biased may express a technical observation that it achieves

worse results, measured by accuracy, on some population relative to others. Conversely, it may express a normative belief that algorithmic measures of success, such as error rates, predictive performance or other technical evaluations should be identical over different populations, at least to some baseline standard or threshold, or that their sample sizes should be proportional, given the utilities of different decisions for the different populations. Others yet may deploy the term to describe a decision process that upholds ideals of equity and justice or that affirms and champions sociocultural differences. In this section, we consider a few different definitions of fairness and analyze what we see as the core definition that is implicit in much of the emerging debate about fairness in algorithmic medicine and that reflects some of the most important and least celebrated historical roots of the concept.

In the most broad sense, a technological artifact is fair if it is not biased or more precisely, if it acts without prejudice or implicit suspicion biases against members of certain groups or categories of people within a society, meaning currently that the individuals associated with those groups do not experience undue negative outcomes with respect to the design decisions. The observation behind this definition is based on intuitive agreement and scholarly consensus that technologically-mediated decision processes are sometimes carried out with prejudice and craving because the technology is an insensitive or fickle observer, relying exclusively on imperfect estimates of underlying causal properties.

2.2. Measuring Fairness

Determining whether an AI Decision Support System (DSS) is "fair" requires measurement. The most common approach is to use a quantitative metric that aggregates information about many individuals or groups and reflects a concept of fairness [18-20]. For example, group fairness metrics define fairness in terms of similar statistical properties between two or more subpopulations. Commonly used group statistical measures include risk difference, risk ratio, or odds ratio. These aggregative measures can be further quantified into fairness metrics that compute a measurable metric, such as Demographic Parity Equality and Equal Opportunity, across different subpopulations. However, quantitative population metrics have been critiqued for:

1. Focusing only on population statistics rather than individual-level decision outcomes; 2. Measuring fairness for specific subgroups rather than assessing the entire population; and 3. Using a single fairness measure can be insufficient — selecting a complication metric that often conflates different user needs.

An alternative approach to measuring fairness is participatory justice, which stresses the importance of gathering subjective information from stakeholders about the intended usage role of a particular AI DSS. Implementing participatory fairness can be

just as challenging and resource intensive as the participatory approach used to heuristically assess the manner of development of the AI DSS. Participatory modelling fairness metrics require stakeholder input on standards that are influenced by how the AI DSS is intended to be deployed, and are then aggregated into equity functions. Such metrics can be applied for decision audits but lack clearly formulated, formal implementations and testing procedures for enacting these measurement activities. If the intended role of the AI project transcends the local context, then fairness measures that focus on the entire population, rather than on subpopulation groups with well-known simplicity biases attracted by inquiry, need to be enacted.

2.3. Challenges in Achieving Fairness

Achieving fairness in AI systems, and in predicting models more generally, is difficult or even impossible and it is important to understand these difficulties [19,21-22]. The theoretical impossibility results show that if fairness is defined in similarity to error parity but by considering conservative sets of protected attributes, then fairness with respect to such a conservative set is not achievable, or it is achievable only under unrealistic assumptions, such as zero conditional error of the model. These results place serious restriction on the types of predictive models for which achievable fairness is guaranteed. In the other direction, one could define fairness differently. For example, error parity can be relaxed to be similarity to error parity [11,23-25]. However, as shown, it is possible which leads to much less interpretable normalized measure of fairness than counterfactual fairness and the problem of counterfactual fairness is based on error parity.

First, such disparities in error may well be acceptable as long as they are not too large. Model developers tend to place more meaningful constraints on fairness, which rather than purely a postulate to be followed automatically, will tend to be a direct reflection of the very aims of machine learning. Second, the pitfalls inducing such disparities are not merely artifacts of the data or evaluations used. AI results in discrimination and inequality when the actual prediction model is used to affect lives and fortunes. It is for this naive reason that one could ask whether counterfactual fairness of the model is attainable or ever would be aimed at.

3. Accountability in AI

3.1. The challenge of accountability arises particularly acutely from the use of algorithms or AI to support, aid with, or partially replace human decision-making [26-28]. It is not a new issue and relates to questions around delegation of authority, changes in the responsibility of the primary decision-maker, potential erosion of the relevance of human judgment, and diminished culpability in the event that the decision or action against which someone is accountable leads to an event of moral or legal

significance. When accountability is pursued too much, or incorrectly, it can inhibit the development and use of important technology partnerships and lead to uninformed strategies for use. Accountability can take several forms. Legal accountability arises from connections to the legal system, can appear as direct or indirect power over a decision-maker, and for judgments and actions can rely on frameworks of tort, criminal, or regulatory liability. Moral accountability involves accountability and moral judgments made by third parties, and may or may not relate to formal legal accounting. Delegated accountability can arise to allocate to a given actor responsibility for the actions of agents acting on behalf of others. Technical accountabilities can take forms such as explainability or auditability that create requirements for algorithmic transparency. These meanings describe a shared domain of accountability concern but have separate areas of focus. 3.2. Accountability Mechanisms AI accountability is often assumed to be primarily a matter of legal responsibility: either applications or users will be legally accountable for the consequences of AI judgments. This is unlikely to be the solution strategy for any AI application or for any specific case involving AI systems. Consider both direct and indirect liability in tort or criminal law terms: direct liability may not be possible because AI liability recognition is likely to be limited; core elements of intent or negligence might be impossible to meet. Indirect liability should be limited: already overused doctrines of vicarious liability could be obscured further if the company uses AI and makes decisions based on AI judgments.

3.1. Understanding Accountability

Accountability concerns in relation to AI in healthcare stem from the dynamic and entangled socio-technical systems at play. Due to the increasing automation of services and tasks that augment or replace human decision-making, questions of accountability become complicated [29-32]. Instead of reconfiguring existing systems where individuals are already responsible for the decisions of others, the development and application of AI systems reconfigures the systems of accountability by introducing new actors and tools. We stop being account-givers towards the designated actors and become account-givers towards a black box that produces outputs we rely on. As technologists and decision-makers, AI systems designers and manufacturers are included among the group of people whom accountability is distributed between. At the same time, there is also a risk that technologists and decision-makers try to monopolize accountability, directing the burden of responses back towards the individuals impacted.

Accountability is multifaceted and this section does not seek to provide an exhaustive account of the philosophies surrounding it. Broadly, accountability entails a form of social relation in which some agents are responsible for some set of other agents,

arbiters, or evaluators [31,33-35]. The latter have a normative expectation for assessing whether those actions have met some agreed-upon criteria. The terms of this relation can change, however, meaning that a person can be accountable to another person and later switch, without relinquishing their previous accountability. In terms of the actions being evaluated, those who are accountable can be expected to answer for their actions' outcomes, processes by which those actions reach outcomes, or even the capacity in which the person acts. Some philosophers argue that responsibility confers an obligation to answer in the event of harms in the form of blame or punishment, while others hold that to be accountable is to explain or justify our actions.

3.2. Accountability Mechanisms

Accountability's importance depends on the potential scope of an AI system's impact. Technologies with little impact, such as small automations or tools, tend to be used with little governance [36-38]. In health and medicine, clinical decision support systems, diagnostic models, and all systems at the core of care may actually be designed for low scrutiny of decisions, either directly or indirectly. However, the actual mechanism of accountability is often outside the design process per se. We also see a border case with powerful research systems, which should be exposed to scrutiny for social benefit but are usually treated like any other piece of research. Conducting research can be an accountability mechanism, but reporting standards might be lacking or even missing entirely. For impactful applied systems, such as those that are core to daily clinical activity, concluded processes materialize: the workflows and tools to ensure that models are validated and tested in the field and that their shift might alter processes that ensure that we don't make lousy decisions become part of daily work.

Accountability is described as a way to make sure that the developers or deployers of a system are in a position to be adversely impacted by their decisions. In organizational theory, structures have developed to assess at scale accountability across a group: transparency, trustworthy tracks, and common incentives are aggregate mechanisms often nourished by scandals. Trust is often mentioned but rarely defined, so here we discuss four potential definitions. AI systems in health or medicine do not need specialized trust. The traditional one is coming from experts and users, such as physician colleagues. This trust is used for lots of systems, and more profoundly for higher-impact or riskier systems, so it is very specific to each system, but almost without exceptions has to be crossed for higher stakes or critical use.

3.3. Case Studies on Accountability

Accountability is required within the healthcare space long before issues of blame and retribution arise. Instead, it is regarded as a key mechanism in the regulatory toolbox for establishing the necessary institutional foundations that govern cutting-edge

technologies. From science to its applications in healthcare, AI and ML involve actors from diverse backgrounds working within complex regulatory, ethical and commercial environments with third-party and proxy interests. AI and ML are not simply a tool or technology used by these actors; it often establishes a proximity between the outputs of the technology and the goals of a company, gauged in terms of profits, making the technology liable to the level of public trust in these organizations. Trust in proximate actors, and their motives, is integral to an organization securing social license to operate, which may be disrupted by the output of an AI or ML system that is unfair or inequitable to stakeholders, resulting in reputational harm to the organization. Additionally, the used datasets carry significant moral weight, and enabling the accountability of proximate actors involved in this aspect can act as a form of privacy protection. The responsible use of AI in healthcare cannot simply be made upon regulation of the final AI product, but must also address the accountability of these proximate actors involved in the entire AI lifecycle [1,39-42].

In a survey regarding data science governance, a majority of respondents considered the assurances provided by platform companies offering AI-related services to be inadequate. Similar issues have been raised regarding accountability of large development teams within a tech company responsible for releasing AI-driven technologies. This is even more pronounced with the rise of "AI as a service" wherein model development and deployment is offered as a digitized service to third-party organizations. The analysis has pointed out that this is undermined by a lack of mechanisms to cope with the Greybox problem associated with machine learning, wherein test suites incorporate the variable traits of diverse users of disparate backgrounds, working within different environments. The result is inadequate service provisions that may potentially violate private information or disrupt the reliable operations of a user, possibly having unintended effects at a larger environmental scale.

4. Transparency in AI

4.1. Importance of Transparency

AI has the potential to radically transform healthcare, yet it raises important ethical, legal, and societal issues, many of which warrant further examination before it is deployed at scale. One of the most pressing issues surrounding AI – indeed, for emerging technologies in general – is how to strike the appropriate balance between fostering a culture of technological innovation while ensuring sufficient levels of oversight and accountability. Answering these questions requires serious consideration of issues surrounding AI that are specific to the healthcare context. For example, given the importance that transparency has in other contexts, one of the first questions to be

asked regarding transparency concerns the systems level: how does this affect the already opaque nature of the healthcare system?

In order to have a functioning healthcare system that respects the autonomy of the patient, patients and their families must know what treatments are available and how they work. In order to assess the likely success of a particular treatment and, assuming the patient gives informed consent, weigh the risks, benefits, and chances of success against such metrics as reduced quality of life, expense, and time and energy required, it is important to have knowledge of not only how a given treatment works, but also why it was prescribed. The importance of these considerations is amplified when it comes to AI and healthcare. Not only will patients have a more difficult time understanding and questioning recommendations made by AI, but if implemented inappropriately, AI may create a disconnect between the patient and healthcare provider. Such disconnect could be particularly troublesome in cases when they are already vulnerable.

4.1. Importance of Transparency

Transparency is one of the important characteristics of AI systems that are to be deployed in high stakes domains where poor decisions can cause severe bodily harm or worse. Medical AI provides decision-making support related to patient diagnosis and prognosis, as well as treatment recommendations and other choices. It is also used in determining third-party coverage decisions, including insurance policy decisions. Increasingly, AI tools are being designed to be integrated into the processes of care as tools for decision-makers like clinicians, patients and their families, insurers, and other stakeholders to facilitate their choices and to ensure the best patient care. We believe these systems hold great promise for assisting stakeholders in addressing the challenges they face in providing and receiving the best quality of patient care and improving the efficiency of corresponding healthcare delivery processes. Transparency is central to achieving this promise.

Taking together the basic principles of transparency, it is clear that AI cannot simply be a “black box.” But how transparent should medical AI systems be? With algorithms as sophisticated as modern AI decisions systems, we will be confronted with challenges in meeting strict standards [7,9,10]. Another meaningful form is model transparency, the ability of a human to see how a model works, what kind of patterns it learns and so on. In the medical domain, we as the AI community need to support our clinical allies in developing understandable and interpretable models. These allies are the domain experts, the ones trying to make sense of these models, and together we should find ways to make sense of them, as understanding enables them to trust our work. What is critical to a model’s interpretability in the medical domain is whether it

can produce understandability and how the answer impacts medical decision-making and outcome. In sum, what is at stake is enhancing user's cognitive understanding.

4.2. Techniques for Enhancing Transparency

Machine learning is an indispensable tool for many, if not most, AI systems. It can thus be quite helpful to understand how to make the operation of ML systems more transparent, which in turn helps to make their functioning apparent. Transparency techniques can thus usefully be divided into two parts: tools for making ML systems more transparent, and tools for providing transparency to ML systems. The first reins in the opacity of a system by design, while the second aids in discovering why such a system produced a particular result.

In the first category, we can place the concept of interpretable models, that is, a model that is relatively easy for a human to comprehend. There are a variety of machine learning algorithms that are considered to produce interpretable models as they create explicit decision rules laid out in an easily readable manner, as compared to neural network-based models which are usually seen as black boxes. Interpretable models, however, have their own limitations. Their use is often restricted to small datasets and low-dimensional classification tasks, and there are certain inherent limitations to the predictive performance of such models. Reducing the number of weights in a neural network model is one of the approaches to providing transparency by design while permitting performance increases. Since more weight often correlates with better performance, this may not always be a viable approach. Reliability of results can also be used as a metric for allowing the widest number of people access to decisions made by machine learning tools and taking trust out of an utterly black box model. However, this has to be weighed against the accuracy of predictions.

4.3. Transparency and Public Trust

Nothing in AI and healthcare raises public concerns like an aversion to opaque systems acting in black-box mode. A long history of medical screening tests, although functionally never as powerful as they have been in recent years, brought patients to diagnose serious medical conditions with the problematic potential of overhauling people's life to the point of breaking them apart. Invasive tests to detect someone's HIV, HCV, HPV, or TB serology, among other pathogens, often require the act of looking back for life-challenging pathologies to not only affect the life of the person involved, but also their family, society, and the state. Thus, deviant outcomes can drive an entire set of further actions. What about unexplained errors being done in such sensitive decisions by AI-enhanced systems — thus worsening the patient's condition with wrong or deviant risk stratification to trigger loss of public trust and confidence in the technology?

However delicate the subject may be, exploitative models without validation cannot be simply rejected from using AI and machine learning concepts when in clear disfavor to the negation of an important upside. In fact, black-boxing the model's decisions leaves these results unpredictable and untrustworthy, which in itself constitutes a significant obstacle to the deployment of machine learning-enhanced applications. Public and regulatory trust in algorithmic decisions is particularly important in sectors such as healthcare, the law, and public service, suggesting a trade-off between accuracy and explainability. Interpretability matters not just because it can alter the choice of which algorithm to use but also because it plays an important role in determinations of trust in algorithmic decision-making. Such a dilemma accounts as one intrinsic property of complex AI-enhanced systems with a growing complexity, or deep learning models, such as those based on neural networks. Because of the growing number of validations required by regulatory authorities, efforts have been made, for instance in the area of medical imaging, with additional constraints being imposed to improve interpretability. Indeed, no matter how romantic an idea reinvention of a development process might sound, until there are clear answers for consequential problems touching upon fully consequential existent questions, public trust and confidence matter most.

5. Regulatory Frameworks in Healthcare

The need for a regulatory framework for artificial intelligence (AI) in the field of healthcare is critical. As AI grows in importance and inevitably becomes more integrated into the field of healthcare, regulatory bodies will need to set forth guidelines and requirements to ensure that these solutions are safe for patients. The frameworks focus heavily on the fact that AI is a "black box" with complex machine learning (ML) algorithms, and thus warrants additional scrutiny before deployment in a healthcare setting. As 99% of the drugs that go through the approval process are unsuccessful, this prolonged validation process becomes even worse for developers of AI-based solutions, as the time from inception to use in the clinical setting can take up to a decade or more.

Additionally, the framework document includes a comprehensive framework for telemedicine devices that include AI, which further delays the approval process. Consequently, the complex development pipeline and approval process lead to a situation where AI developers may choose to forgo regulatory approval completely and deploy their product directly to the end user. This unfortunate reality of a no-regulatory environment complicates the reality of developing trust in AI and ML tools. Similar recommendations about the need for post-market real-world evidence generation and device performance evaluation are made. In addition, the need for transparency, accountability, reliability, and cybersecurity are also key recommendations. AI

development is a new and dynamic field; thus, the regulatory frameworks are not static and are expected to evolve as new devices and technologies come into play.

5.1. FDA Guidelines for AI in Healthcare

The FDA is responsible for protecting the public health by regulating the safety and efficacy of products. The Center for Devices and Radiological Health oversees medical devices, a category that is growing rapidly to include standalone software and software in combination with drugs. In 1996, the FDA issued a guidance document describing how the FDA would evaluate software. In 2019, the FDA announced the creation of the Digital Health Center of Excellence, which focuses on accelerating digital health innovations, especially related to software. AI/ML-based software constitutes a subset of medical device software. The FDA has not targeted pathology software as a distinct type, but the guidance documents apply specifically to software that is standalone or used with a physical medical device, including some types of imaging algorithms. By these criteria, the digital pathology algorithms for input used in machine learning are not exempt from regulatory review. It is of note that there are no quality system requirements about design control or postmarket regulation requirements for radiation control. Because it constitutes the highest risk, any approval will not happen until all regulatory hurdles are cleared.

The FDA has issued two guidance documents specifically addressing the regulation of AI-based software in healthcare. In a 2021 document, the FDA outlined a framework for approaching software algorithms that would rely on some in-market real-world review for ongoing updating and continuous learning. The aim is to make approval and regulation of AI devices easier and faster for devices that would change and evolve continually. A second and more general document was issued in 2019 that introduced transparency, explainability, accuracy, robustness, and cybersecurity as important pillars of regulatory review. A significant portion of the FDA's regulatory framework for algorithms was based on the real-world experience with mammography screening.

5.2. EMA Regulations and Standards

The European Medicines Agency (EMA) is a decentralized agency of the European Union (EU) with the mission of protecting and promoting human and animal health, identifying and characterizing health threats, and minimizing their impact with a comprehensive and coordinated approach. To achieve this, the EMA is responsible for the scientific evaluation of medicines developed by pharmaceutical companies for use in the EU.

The EMA is not responsible for the regulation of medical devices. Identification and characterization of health treatment needs in the European Union for both medical

devices and in vitro diagnostic devices are the responsibility of the Medical Device and In Vitro Diagnostic Device Regulation. The European Union Regulation on Medical Devices provides the regulatory framework for the conduct of pre-authorization clinical evaluation required for the authorization of clinical investigations with these products. Such clinical investigations are sponsored and conducted by companies and institutions with the appropriate expertise and resources according to Good Clinical Practice in accordance with the clinical investigation plan, approved by an ethics board and the relevant competent authority.

The clinical evaluation of AI-based technologies applied to medicine in the EU is performed by authorities in the member states of the European Union. The purpose of the clinical evaluation is to confirm the claimed medical benefit, establish that the general safety and performance requirements of the device have been fulfilled, and identify any remaining risks and contraindications to the use of the device. The manufacturer and the authority should verify that the information for use is compatible with the medical benefit claimed and ensures safe use of the device.

5.3. Comparative Analysis of FDA and EMA

FDA documentation considers an AI/ML as a device, while EMA regulation documents do not discuss AI/ML as such. The difference in general definitions of AI/ML is interesting. FDA documentation refers to the variety of applications as "algorithms", and considers AI as a “device”, holding the position that "an algorithm is a medical device". FDA considers that, “When software is intended for general and unspecified use as an adjunct to a medical device, it is considered a component of the device.” FDA also states that, "When a medical device contains software that meets the definition of an intended use regulated under 21 CFR 807 or 21 CFR 814, the software is subject to premarket submission and clearance/approval".

EMA regulation documents do not discuss AI/ML as such. Both FDA and EMA do not discuss governing AI/ML software systems as a “component” of a larger software system, which is the case for many AI/ML software systems used in healthcare and life-sciences. A few AI/ML software systems are standalone devices governed as a “device” by the FD&C Act in USA or as devices under the EU MDR. For example, for the specific product classification including data and “software as a service” clinical decision support product that uses the AI/ML software technique without any “real-time” connection to any upstream healthcare and life science related device system networking environment.

While regulations in USA and EU might appear to be onerous, and overly complex for startup companies, especially AI software developers not familiar with regulatory approval and validation expected for their product, it does serve the purpose of

instilling confidence among the users of AI/ML and other systems and is an important pre-requisite for adoption of such technology. Adoption is likely to be more expedient if the AI/ML systems have been approved according to regulatory standards and validation by a government body responsible for citizen healthcare protection.

6. Environmental Policy and AI

How should environmental policy respond to artificial intelligence capabilities? This question has already arisen in a few environmental areas. The processes by which the executive branch actors adopt new regulatory initiatives include relevant mechanisms to shape such responses. Further below we survey how new interagency initiatives may be used to shape those responses.

Few agencies have as prominent a role in earth and ocean science as the National Oceanic and Atmospheric Administration. NOAA's underlying statutory frameworks are broad grants of authority to monitor climate, weather, ocean, and other aspects of both terrestrial and marine environment. Through a number of specific statutes NOAA is charged with responsibilities ranging from air quality monitoring, charting the nation's and the globe's oceans to managing coastal zone information and aquaculture to ensuring the safety of fishery products. Within the agency, the National Environmental Satellite, Data, and Information Service operates the nation's environmental satellite system and provides global environmental data. AI and machine learning increasingly support these remote sensing missions. Additionally, NMFS has a longstanding role in monitoring commercial fishing and seafood processing using a specialized AI known as catch share. Outsourcing these roles to the private sector has received renewed interest recently due to the expansion of AI capabilities like the ability to better evaluate tradeoffs associated with contractual as opposed to regulatory long term fishery management decisions.

6.1. NOAA's Role in Environmental AI

Environmental AI is hosting the promising partnership of the National Oceanic and Atmospheric Administration and the federal policy on the regulation of Artificial Intelligence. From its very inception as the Executive Order on Maintaining American Leadership in Artificial Intelligence, Agencies were mandated to create their own policies, and further promote the National AI Strategy. The Administration has recently proposed a new opportunity to foster support of nationwide AI's research in the workshop on Reinvesting in AI Research: Perspectives from the Public, Private, and Non-Profit Sectors. Therefore, national and federal policy agencies have been actively developing a position for that Policy Sector in establishing guidelines and regulations where AI will be applied, employed, and used.

The National Oceanic and Atmospheric Administration has developed its Environmental AI Strategic Plan to describe how Environmental AI will help fulfill its mission to better understand and predict changes in Earth's environment, from the depths of the ocean to the surface of the sun, and share that knowledge and information with others. Environmental AI will serve the long-term goal considering that AI is a tool to amplify the extraordinary work of NOAA's employees and partners. AI leads to priority objectives that guide efforts in Environmental AI and will provide direction and ensure focus as they invest in Environmental AI over the next five years and beyond. The Environmental AI objectives are to promote research on new and improved Environmental AI methods, tools, and technologies to meet mission needs; Increase the impact of Environmental AI within operations, products, and services to better serve our Nation; Promote the transfer of successful Environmental AI capabilities to operational use; and Lead and participate in cross-agency, national, and international collaboration on Environmental AI.

6.2. NASA's Policy on AI Applications

NASA's charter requires them to "pursue the widest practicable and appropriate dissemination and use of information and technological developments ... to enhance economic, environmental, and social benefits to the United States", including "the prevention of loss of life and property from natural disasters." AI applications and use, to include Machine Learning techniques, are difficulties for which there is an immediate need for agency interval collaboration and collaboration with other governmental entities and the private sector. NASA endeavors to provide accessible technical evidence, evaluation, guidance, and knowledge useful for society's understanding of the potential uses of AI and the benefits, consequences, and influences of its development. Policies governing the application of AI tools, Machine Learning, by NASA include ensuring that: AI design and applications meet the mission needs of the Agency and its partners; stakeholder and community needs are assessed and AI activities are evaluated across disciplines. Agency AI and Machine Learning capabilities are cross-domain technologies, broadly utilized across NASA's mission areas. A primary focus for NASA AI research is on critical technical constraints that arise from operating in demanding real-world environments such as space exploration and use, aviation safety and use of the increased airspace, safety or surface operations. The primary aim is to develop advanced technology and systems that enhance safety and reliability by integrating AI/Machine Learning capabilities into existing and future systems. Secondarily, to prepare for the future by developing technologies to enable safe use of higher levels of autonomy in transportation systems and to ensure the success of AI-enabled NASA missions so that they can adapt readily to uncertain and unanticipated changes that affect mission success.

6.3. Interagency Collaboration and AI

Like other Federal agencies have expressed interest in incorporating LLMs and other AI platforms in their operations. You might think that those are just competing for the latest technological glitz. While that desire for advancement may be present to a degree, there are good reasons to investigate these systems from an interagency perspective. Each has a very different user base, different competencies, and potentially–testbed data sets. The entire Federal enterprise will benefit from collaborative exploration in this space.

This is not a new concept. Billions of dollars have been spent on interagency collaboration for decades, especially in the area of defense and intelligence, but also in cloud services and general infrastructure support. The mission-driven, safety-conscious applications of AI in the analysis of the decision spaces we face in many of the areas in which service agencies have been long involved is ripe for broader, deeper investigation. Crazy ideas, like a battle management system designed to enhance Crisis Action Team functioning in an information overload environment, shouldn't be the province only of research comedy shows. Tackling integration of FedGov use of LLMs in expert taxonomies, "almost everything else sensing," and focusing optics for long-term fusion decisions deserve far more than their current momentary glances.

All of our communities would benefit from a coordinated position statement identifying what capabilities may be useful for which agencies in interacting with the agencies' populations. Do complaints of insufficient presence in its AI-biochar rulemaking wish list imply that both it and could benefit from positioning guidance involving the tinier than current 10-stack?

7. Responsible AI

In the face of increasing algorithmic influence on key aspects of our lives, discussions of how to implement AI responsibly have gained momentum. In this section, we focus on AI for Biomedical and Healthcare Research and Targeted Delivery, and define Responsible AI as considering both the AI tools used and the desired outcomes of AI applications. Specifically, our definition underscores the need to align trade-off decisions in AI implementation with (1) stakeholder needs, (2) consideration of intersectional marginalized communities, (3) collective values, and (4) resource equity. AI algorithms might require careful design and implementation to achieve equitable and just societal benefits, while withholding their use when it could lead to inequitable trade-offs. Questions, such as “What social or cultural values does this algorithm promote?” or “Which dimension of wellbeing is being improved, for whom, and at what potential cost?” can help to critically assess AI in terms of Responsible AI. We frame Responsible AI as similar to the Ethical and Responsible Research and

Innovation discourse, but emphasize the unique considerations of the algorithmic context.

Recent years have seen an increase in both technical and thematic proposals for how to implement AI responsibly. Many thematic proposals echo the historical developments within HCI concerning ethics and values in Design, calling for Algorithmic Impact Assessments to ensure that AI tools will fulfill their promises, and dedicate time and resources for Design Justice prior to deployment. Many also arise from criticisms of problems endemic in the design and use of AI tools, calling for Bias, Fairness, and Equity in Algorithmic practices. The recent frameworks in the UK are examples of responsible implementation efforts. Technical proposals for Responsible AI often propose technical methods to reduce Bias in learning, such as Fairness-Aware Machine Learning or methods to introduce Explainability as a part of AI tool design – sensitive to all stakeholders involved in the decision making process of the algorithms.

7.1. Defining Responsible AI

Responsibility appears to have a natural connection to what it means to be human. We make decisions, we act, we are blamed or praised for our choices and their consequences. In this regard, we can say that we act responsibly when we freely choose and perform an action having in mind that action's value and acknowledging the relevance of its consequences. Minimally, a person who acts responsibly must possess the freedom to act, the capability of deliberating about possible actions and their consequences, and the capacity to empathize with those affected by the action. Especially in the age of AI and the use of models that operate independently of human judgment, the notion of responsible action raises questions about the relevance of traditional criteria on human decision-making. This is particularly evident in discussions about AI and machine learning bias or fairness: who is responsible for biased or unfair results produced by algorithms? The designers, the developers, the providers, the users of those algorithms? Or, since algorithms are allowed to act independently, are the algorithms themselves morally responsible for biased or unfair outcomes?

The use of the term 'responsible AI' should not be understood along this line. It is not about assigning responsibility to the algorithms themselves. Rather, the use of 'responsible AI' is a shortcut to invoking the idea of responsibility with regard to the deployment and application of AI and ML in the real world. More generally, we can refer to Responsible AI as a set of principles or guidelines about how to analyze, design, plan, and deploy AI technologies in ways that foster and not hinder people's wellbeing, meaning and purpose, social belonging, agency, human rights and dignity, and flourishing as individuals and in their community.

7.2. Principles of Responsible AI

Several AI principles have been issued for public commentary or approval, by public and private entities, and there are some commonalities among them. The first principle consists of promoting the common good. AI technologies and their associated data should promote the public interest and guaranteed rights. Computer algorithms should not contribute to existing or perpetuating inequalities or discrimination. The second principle of responsible AI is transparency, which consists of openness about data collection, use, and management. Users should be informed of how, why, and when an AI system is being used in order to understand its limitations. Lack of transparency is a barrier to accountability. The third principle is about privacy, which should be actively defended. Privacy protection should be default, and organizations must respect individual privacy as an ethical consideration, for its own sake, and any reasonable requests to erase data should be done in a timely manner. The fourth principle consists of AI systems being socially aware. Systems must be built to operate in a diverse social environment, and training and design data sets must be representative of end-users. The fifth principle is about accountability, which demands that AI developments should go through a rigorous exam and approval process and that individuals be appointed for the management of specific impact-related activities.

The sixth principle is altruism. Creativity, the arts, and care-giving activities should be respected and enhanced, and noneconomic interests related to the use of AI technologies should be preserved. The seventh principle brings up the matter of obsolescence, suggesting that preservation from excessive and unintended obsolescence of the cognitive and non-cognitive capacities of individuals, as well as the capacities of regulatory agents like governments and organizations, should be assured. The eighth principle refers to AI steering. AI systems should not steer undue decision power away from individuals towards technology.

7.3. Implementation Strategies for Responsible AI

Most organizations recognize the importance of being transparent, fair, and accountable, but they may be uncertain about how to enact the principles of responsible AI. Specialized teams within organizations can be charged with experimenting with and instilling generous use, shaping use policies and guidelines, and auditing algorithmic systems and their outcomes. Organizations might install AI review boards and interdisciplinary ethics committees to scrutinize ambitious projects involving predictive algorithms and algorithmic systems to ensure that those systems approximate good behavior. Organizations might also create expert outsourcing agencies to audit algorithms. Training, education, and research are important to various aspects of AI development and deployment, as is legislative and legal oversight.

For organizations developing and using AI, research exploring self-regulation in addition to the traditional, regulated management of health-related technological innovation could provide important insights into best practices for adopting and implementing responsible AI. Industry associations might adopt consensus ethical guidelines or ethical credentialing of member organizations. Contracting vendors or stipulating gain-sharing agreements with certain partners in the AI ecosystem might help encourage responsible AI pursuit for profit. Additional discussion surrounding core investment of algorithms used for structured prediction and clinical decision support systems in healthcare might help ensure that the principles of responsible AI are paramount. Lastly, examination of how existing and proposed legal vehicles governing products, services, and organizations—which fall short of addressing AI-specific issues—might lend insight into how organizations and systems contribute to responsible AI guidelines, principles, and implementation strategies.

8. Alignment with Public Good

The rise of AI applications in healthcare has exposed the pressures that the profit motive applies to the development and dissemination of potentially important new technologies. The motivation for rapid commercialization of AI-based tools is not inherently bad and, when properly aligned with public good, economic incentives can in fact accelerate technological progress. But the deployment of some AI applications in healthcare also has the potential to cause serious, long-lasting damage to individual and community well-being, and to undermine social equity and justice, unnecessarily inflating burgeoning healthcare costs. Therefore, it seems crucially important to consider how public good can be factored into the AI development pipeline—from initial conception, to funding decisions, to design choices, to validation and testing, to eventual deployment.

It is perhaps useful to begin with a definition of public good. The term describes certain types of commodities or assets that bring benefits broadly across society. Depending on the context in which the term is being used, public goods may be broadly defined to encompass services furnishing the most basic determinants of human welfare and dignity on the one hand, or more narrowly defined to include things like water supplies and air quality on the other. There is a common challenging aspect for all public goods, which is that there may be little or no market incentive for companies or individuals to invest in their production. In other words, public goods are underprovided by free markets because there is no direct market demand that ties investment in their production with the monetization of returns on that investment. While healthcare is not literally a true public good as described by economists, it nonetheless shares important characteristics with true public goods and therefore be subjected to a similar logic about the incentives needed to optimize its quality and

distribution. Benefiting from economies of scale, and accrual of secondary external benefits, high-quality healthcare can generate positive spillover effects that incentivize public investment.

8.1. Understanding Public Good

M. K. As stated previously, AI is not to be used blindly; it must be in service of a goal articulated by people, who know the context and nuances of the issues for which AI is applied. AI can be used towards many kinds of ends, but the ends we advocate for are those that align with the message of medicine, which is aimed at decreasing suffering in a complete and thorough way. These goals serve to frame ethical considerations that will be discussed throughout this chapter and the book. What do we mean by public good? It seems to be a simple term. We may construe this to mean anything that advances the greater good or the greatest amount of pleasure for the greatest number. But we have four words in the phrase – public, good, advance, society. As with all phrases that seem direct and uncomplicated, when contemplating them deeply, we find a richness of meaning and nuance that evokes further meditation and analysis. To take the last term first, we focus on society above all because we are animals of the collective, and we are fundamentally dependent upon our people for our survival. While there might be individuals who seek self-gratification above all else, the greatest satisfaction comes from being part of a group. As renowned psychologist found, intrinsic motivation is enhanced by social relatedness. Similarly, psychologist identified flow, one of the most pleasurable of all psychological states, as being one where we are absorbed in activities in service of others. Above all, our histories and our anthropology demonstrate that evolution has favored social cooperation and development and that our survival depends, ultimately, on the group.

8.2. AI's Role in Advancing Public Good

Global challenges, such as climate change, pandemics, and geopolitical conflicts, often exceed the capacity of national governments and multilateral organizations. As the urgency of response necessitates mobilizing new, coordinated individuals, organizations, and governments at the local, national, or global scale, technologies are required to work towards the resolution of identified issues at a scale and speed that was not previously possible. If technology alone is not sufficient to meet the social demand for global public goods addressed by an Effective Altruism or Global Priorities approach, at least it is necessary. AI has the potential to shift the cost-benefit analysis in favor of investing efforts and resources into responding to challenges to humanity’s well-being and survival, such as natural disasters and violent or nonviolent instability. These goals vary from providing coordination capacity during crises to increasing longevity and well-being, reducing suffering, increasing suffering, and

improving financial stability. Additionally, AI can optimize tasks that are veritably monotonous or rely on time-consuming processes.

Nothing, of course, ensures that organizations will dedicate themselves to the production of public goods; motives, emotional inclination, or ethical commitment are not uniform and can change over time. Companies may have strong incentives to undertake initiatives that generate an aura of respectability and may be put under social and market pressure to enhance the public benefits from what they do. For these reasons, many see the role of government as critical. Through its mechanisms, internal and external, it must push the economy and society to adopt and institute measures that allow the use of the most advanced techniques, as they are what will allow the achievement of the Society 5.0 goals. Society 5.0 can only be achieved if both private persons and companies opt for advanced self-governing solutions through the responsible use of AI.

8.3. Evaluating AI Impact on Society

This section summarizes consider some approaches that can be engaged to support evaluation and demonstrate the type of understanding and distillation of commonly understood evaluation truths, facts, activities, and observable measures that should be instantiated when Artificial Intelligence (AI) research concerning its realization and deployment in support of public good. Evaluating the impact and benefits of AI on society is a big and important question that might rightly seem to ask for the invention of a new form of digital econometrics. It encourages collecting actual test data, creating surrogate models and new deployment systems benefiting from considered understanding, and delivering report data or metadata on attribution and causation concerning the predictive ability of the developed model concerning real-world responses induced by actions.

That said, the alternatives to using recognized econometric methods would be the creation of general data model priors making many types of evaluation studies private and of questionable utility. To use crowdsourcing systems to engage rapid evaluation by asking crowds of different people to describe what differences the deployment of a new AI may make to their own or others' lives and net benefits achieved seems more directly usable and wonty by an evaluation stakeholder who stands to be impacted by a deployed AI's actions rather than by a popular or commercially competitive impetus to dream big. The difficulty with any crowdsourced approach to model desired outcome states and scoring is that, without some filtering, it could serve to amplify the negative and the hurtful without truly evaluating why an AI might deliver worse outcome utilities but are called out for special consideration.

A more ground-up actionable approach might be to incentivize stability-fans – those people whose actions tend to be the most predictive of long and short term public utility as usually developed by measured market incentives – to second-guess and brainstorm hesitant statements calling for special consideration. This directs evaluative crowdsourcing to the people in the best position to sort and score crowd data for identified retroreflectivity public consequences discovered directly through scaling interactions with identified deployed AIs.

9. Case Studies

In the previous sections, we discussed different issues related to the Ethical, Legal, and Societal Considerations in AI and Healthcare including challenges to socially valuable AI. We argued that the previous policy and governmental frameworks regarding AI and healthcare may or may not exist but are inadequate by not providing appropriate guidance for such an important mission. With respect to the particular societal domain of the question of application of AI in healthcare, it is important to discuss case studies before coming to conclusions. The purpose is to offer brief yet to the point examples of such issues on the paper: with very little AI, no governmental framework, nor strong policies, decision-making, issues related to safety, and inequality, or with / without inclusion of human and AI agents jointly making health-related decisions. In the rest of the section, we offer three case-studies, brief discussions of AI applications in healthcare, and other sectors. The idea of selecting such case studies with the applications of AI other than healthcare is to highlight the lessons learned in the other important societal sectors such as public policy and environmental management. Why these sectors? First, AI implementation decisions are confronted with ethical dilemmas; and more importantly, public safety is at stake. Secondly, there is a sectoral overlap in environmental health. Thirdly, for public policy, social justice and inequality issues arise such as lack of expertise in the new technology. Hence, the deeply moving ethical, legal, societal dilemmas may be largely similar to make policy recommendations in other sectors.

9.1. Case Study 1: AI in Healthcare

Artificial Intelligence (AI) is increasingly becoming an integral part of healthcare systems, defining the ways in which professionals and patients interact. Interest in AI has steadily grown, especially for use in predicting patient outcomes, clinical and operational decision support, enabling remote patient management, fraud detection, drug discovery, and disease biomarker discovery. Despite several advantages accompanying the increased use of AI in these sectors, its application is accompanied by security, ethical, and legal concerns which require serious deliberation. Many of the discussed algorithms are not easily interpretable, which can present problems in

application to patients. In addition to this, patient data privacy concerns arise from the requirements of large scale data access. At present, there are no specific guidelines or laws to ensure that the above systems uphold data privacy concerns or allow impartial monitoring.

AI and machine learning are playing an increasingly beneficial role in various aspects of care delivery: from improving an individual's access to healthcare to improving the clinical performance of diagnostic and therapeutic decision-making. These capabilities are being applied to early detection of mortality or adverse events along a patient's entire journey of care. AI is also playing a role outside the walls of clinical practice in more effectively managing population health via predictive algorithms that can stratify risk. Increasingly, AI is being harnessed to predict variation in patterns of demand, using risks between demand and capacity to inform healthcare resource allocation decisions. As healthcare organizations are being provided with increasingly granular, real-time data on utilization and outcomes, they are utilizing healthcare data and AI to create feedback loops for sub-groups of patients: designing, implementing, and iteratively refining clinical protocols. Empirical evidence and lessons learned from the initial development and implementation of these models suggest multiple functional and technical requirements to derive high performance algorithms across these problem areas.

9.2. Case Study 2: AI in Environmental Management

The contributions of AI in environmental management are becoming more and more evident, particularly in fields where the automation of large-scale data analyses is made possible by AI's computational advantages. As such, AI techniques have been successfully applied in weather forecasting, climate modeling, environmental monitoring, disaster management, and resource management; they have supplemented traditional scientific approaches to deliver results more quickly, or have investigated problems too large and complex to be manageable without these techniques. With traditional approaches for analysis of complex environmental systems taking months and even years of time, the fast turn-around provided by AI techniques makes them an attractive option. Earth system science enables us to conceptualize and study the complex cause-effect mechanisms of the Earth's systems by which human activity has been affecting global climate. However, concerns have been raised about allowing AI to control important aspects of climate engineering, as they act independently from human judgment and reflect a regime of control distinct from traditional environmental management. Environmental management is an applied science that seeks to improve the relationship between human activity and the environment. It formulates proposals to cope with negative environmental phenomena and, in this regard, it supports and enriches environmental policies. Environmental policies may be regarded as superior

to both environmental management and traditional scientists of the Earth system. Environmental policies define the objectives and regulations under which both environmental management and Earth system science operate, in order to exercise and optimize the political decision-making power over the human-environment relationship.

9.3. Case Study 3: AI in Public Policy

The third case study concerns AI-based prediction research involving U.S. courts, criminal justice, social services, and public assistance agencies, and its influence on certain decisions and policies as they relate to reductions in child maltreatment, foster care placement, and foster care placement reentry rates. This subsection ends with a discussion of related ethical and other issues emergent from such research, and of the need for policy decisions made on the basis of AI-enabled decision-assist tools and software, such as whether to remove or, if removed, whether to reinstate a child, during the pandemic. More generally, we also address the ethical, legal, and social issues associated with predictive policing as well as its related sociotechnical aspects.

Predictive algorithms in this space draw on various data sources: past child maltreatment fatalities and reports that did not end up being categorized as fraud; transgressor proclivities as shown in behavioral, socioeconomic, and medical histories; family histories—especially embedded stressors such as parental mental illness or substance abuse disorders; and current domestic straightened material circumstances, including economic downturn. Algorithms making use of these data then run predictive models and generate output typically designed to flag, score, or rank families and individuals, localities or regions for preventative services and interventions prior to negative child welfare events. The aim is to guide child welfare agencies in prioritizing actions based on informed risk assessments. A key assumption, stemming from various studies supporting these algorithms, is that, like predicted reoffending in criminal justice, the probability of recidivism increases based on the volume of earlier maltreatment events, family risk factors, and the severity of each earlier event.

10. Future Directions

AI is already affecting every aspect of our lives, including healthcare. Within the next couple of decades, AI will help us determine the best treatments for disease, provide real-time risk profiles for disease based on social media and other data, and improve the management of chronic diseases. AI will analyze our experiences to understand how drugs are experienced in the real world, provide real time feedback on how clinical trials need to be adjusted to capture drug effects better, and even help identify various life stressors, imbalance of social networks, and the need for community services to provide holistic health support. AI is not just about identifying diseases and

predicting outcomes better but will also help all of us lead healthy lives by understanding us better. Some of these trends have already begun with the use of sensors, social media data, and big data. New advances in AI such as its ability to intelligently handle missing data, utilize multiple different data types efficiently, and help uncover new disease-causing mechanisms will enable the successful realization of these trends.

However, with all these potentials with the future of AI for bettering our health and healthcare, many pertinent ethical questions continue to remain and be raised. What rules will govern the continued accelerating collection of these massive amounts of data about our health and the analytics that could violate privacy and affect our mental and physical health? What guidelines will become standards for who owns these data and what corporate or government abuse of this ownership would result in consequences? For the types of models we create and the findings gleaned from them, how will bias and misuse be regulated? How will AI be incorporated into clinical practices in a way that augments and does not supplant humane treatment of patients? How can we ensure that the systems we build result in better health for all people in a just and equitable manner? What does it mean to be human in an age where coming into contact with AI has become common? These are a few questions that the emergence of AI in healthcare brings to the fore.

10.1. Emerging Trends in AI Ethics

Concerns for how to design AI ethically are growing, while both society's deliberative mechanisms and organizations' internal infrastructures are evolving to address this need. On the one hand, in 2023 four of the ten existing key requirements are ethics-related. The Act aims to govern safety and risk and to ensure that AI deployment serves the public good, promoting transparency, fairness, justice, and accountability standards. These principles place an ethical governance obligation on the entities developing and deploying AI, and build upon existing international developments. On the other hand, AI ethics initiatives have merged and diversified within large companies. AI-ethics-dedicated resources are being mobilized: provided with budgets, teams, and expertise, they're decentralizing and sending out the touch-up orders to accidentally or strategically located people.

These company initiatives reference values, standards, and principles from earlier shared documents while modifying them internally. However, this knowledge diffuses in a pyramidal way; it is applied to prioritize guardrail proposals for the most relevant use cases. This allows for larger-scale automated solutions developed in-house or bought on the cheap to be ethically downsized. This asymmetry in the ethical governance of AI design and implementation possibly compromises the societal fabric.

Well-oiled capitalism rests on the foundational synergy between large companies and SMEs. If it is true that the cost of non-compliance is highest for larger companies, these shouldn't forget to cater for the ethical exterior of their supply chain. Should these uneven company ethical outsourcing dynamics consolidate, they could further worsen an already damaged ecosystem, where some small players imbue AI development and deployment within their own products and services with principles and others don't. So, should ethical governance outsourcing dynamics consolidate, a growing asymmetry in the labor market could emerge.

10.2. Potential Challenges Ahead

Future ethical, legal, and societal issues probably to occur in the coming years closely relate to the very novelty of implementing AI technologies. The relevant projects at hand will probably be in areas like the revenue potential of insurance companies by using AI to increase the probability of finding a patient being ill and getting an individual policy; an increase in prioritizing voting for people being healthy for not engaging or financing companies developing novel or emerging papers of little or no possible value; the risk of AI being nothing less than digital miotic agents; the money-pot of patients revealing personal data for scientific research; an increasing reduction of public ministers of health; controversies about the refusal of paying by insurance companies or health services the relevant compensation toward patients suffering from an illness which should have been treated by a machine learning system, effectively being parties of conflicts of interests; the risk of misunderstanding AI's probability-based action or predicting process as deterministic results; possible competition in terms of data possession of worldwide business companies or states in order to provide essential healthcare systems or services; and various expectations concerning priorities in the question of protecting the privacy status of personal data.

Moreover, and with regard to the prior topic, the more essential and urgent tasks seem to be establishing and evaluating national or international administrative or civil laws and a potential liability framework. Other delicate issues seem to arise with the uncertain dynamics of on-the-job-life decisions taken by a non-human acting, proposing, recommending, or deciding authority; and with the obligations of such a machine to give a thorough explanation about its elucidating process in order to trace decision-making and/or prediction errors.

10.3. Recommendations for Policymakers

Recent years have seen tremendous advances in AI capabilities. Such advances have also resulted in a growing range of available AI tools for healthcare that could potentially impact patient care and healthcare operations in useful and positive ways. The decision of choosing to deploy such a tool in practice hinges on the legal and

regulatory environment of the country in which the tool will be used. Historically, most countries have had relatively static laws and regulations intended to protect patients and, in a more general sense, the public at large. These laws and regulations are typically not specific to AI. But with the rapid advancement and adoption of AI technology, and how integrated into society such AI tools may potentially become, existing healthcare laws and regulations may become increasingly out-of-sync with how AI technology can be safely and ethically deployed to optimize its potential value in aiming for physical, mental, and societal health.

In order to address this rapid shift in technology and needs, we believe that countries need a willingness to continually review the legal and regulatory environment regarding AI tools deployed for healthcare. Such regular reviews would allow necessary adaptations to be made to the execution of existing laws and regulations that could potentially stifle the safe and ethical usages of novel AI tools in such sensitive areas. Such regular reviews could also trigger new country laws or global guidelines that would allow a country's regulatory bodies to judiciously develop new or additional laws and regulations that are specific to the features and real-world impact of AI healthcare tools.

11. Conclusion

The healthcare sector is undergoing a transition towards a technologically advanced, data–driven ecosystem designed to support both health and care in a health and predictive maintenance model. Health and care innovation powered by advanced technologies such as Artificial Intelligence have enormous potential to make massive positive impact but at the same time these technologies also bring nurtured ethical, legal, and societal questions regarding their use. The past decade has witnessed an increasing interest among researchers, physicians and faculty, funding agencies, and pharmaceutical companies for exploring the potential of AI–based technologies in the healthcare sector. The healthcare sector offers a rich opportunity where one can apply AI–based innovation and transform frontline services thereby enhance patient safety, satisfaction, and outcomes while lowering the overall cost of care delivery.

We hope this volume acts as a catalyst for catalyzing deep exchange and interaction between AI researchers and healthcare professionals. AI requires an inter–disciplinary approach. It is crucial for AI researchers and developers design algorithms and technology approaches applied in the healthcare space along with their medical expert counterparts as active collaborators at every step of the process. An inter–disciplinary approach will help promote accessible, trustworthy, and high quality algorithms and applications in our healthcare systems. We hope this volume serve as useful guide for AI and healthcare scholars, researchers, and practitioners in understanding the

importance of ethical, legal, and societal considerations in the design, development, implementation, and dissemination of AI–based healthcare solutions.

References

[1] Ratti E, Morrison M, Jakab I. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. BMC Medical Ethics. 2025 May 27;26(1):68.

[2] Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. Royal Society Open Science. 2025 May 14;12(5):241873.

[3] Ntjamba FC, Ashipala DO. Impact on and ethical considerations of artificial intelligence on human healthcare. InAI Technologies and Advancements for Psychological Well-Being and Healthcare 2025 (pp. 1-36). IGI Global.

[4] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.

[5] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.

[6] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.

[7] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.

[8] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.

[9] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.

[10] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.

[11] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.

[12] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.

[13] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.

[14] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.
[15] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.
[16] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.
[17] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.
[18] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).
[19] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.
[20] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[21] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[22] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[23] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[24] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[25] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[26] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[27] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.
[28] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.
[29] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.
[30] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.
[31] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.

[32] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.
[33] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.
[34] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[35] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.
[36] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.
[37] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[38] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[39] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[40] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[41] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[42] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 4: Deep Learning for Medical Imaging Analysis

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Deep Learning in Medical Imaging

The advent of deep learning has significantly transformed the realm of medical imaging analysis, yielding automation techniques that not only rival but, in certain instances, even outperform human experts. This chapter aims to provide a pointed entry point for those embarking on a journey through the large and convoluted field of deep learning applied to medical imaging analysis. The scope of this book is both broad and specific; while we primarily delve into medical imaging problems, we apply advanced imaging analysis for data from any source; within the medical imaging domain, we focus on the core areas of detection, segmentation, classification, and registration of anatomical and pathological structures. Furthermore, we also focus on the methodology of deep learning in terms of its specifics, implementation-related aspects, and challenges. The hoped-for ancillary outcome of this book is to promote the widespread adoption of deep learning within the community.

Deep Learning, especially in its recent incarnations, has proven to be extremely effective for the detection, segmentation, classification, and registration problems in a variety of domains including those with non-image data [1-3]. Hence, it is to be expected that Deep Learning will yield correspondingly effective solutions for a wide range of medical imaging analysis problems in the same manner as it has for other domains. However, such an expectation must be tempered; one reason is that most of these AI-enhanced solutions are still in their evaluation phase; to make an impact on a routine clinical basis, such solutions need to pass several stringent hurdles including validation and necessary regulatory approvals, especially those pertaining to ethics,

transparency, and accountability [2,4,5]. Hence, realistic expectations are key to the acceptance of new tools. The high complexity of deep learning models and the massive data requirements for training them and deploying them safely are other issues that should be borne in mind – data availability and privacy restrictions will also play a role in the adoption of deep learning models.

2. Convolutional Neural Networks (CNNs) in Radiology

1. Introduction Deep learning has been an indispensable tool in medical imaging research, especially in image reconstruction, registration, and segmentation. Deep learning is particularly well-suited for tasks requiring discrimination and classification of images from ground truth labels, which is the basis of supervised deep learning approaches. In this chapter, we discuss applications of convolutional neural networks (CNNs), the most common deep-learning model applied to the above tasks, to the medical image modality used most in clinical practice, radiography, and imaging methods for cancer diagnosis and therapy management, computed tomography (CT) and magnetic resonance imaging (MRI). While deep learning techniques are rapidly expanding to new imaging modalities such as ultrasound and photoacoustic imaging, we limit our discussions here to the applications of CNNs to established clinical imaging modalities, without loss of generality, and demonstrate the unique strengths and weaknesses of deep learning by highlighting key features of established techniques and how deep learning can address these challenges.

2. Convolutional Neural Networks (CNNs) in Radiology 2.1. Architecture of CNNs Deep learning is a class of machine learning algorithms that utilize several stacked processing functions called layers to learn the features embedded in the training data. A deep network can learn features of increasing complexity at successive layers. Convolutional neural networks (CNNs) are a class of deep learning methods that are designed to process the high dimensionality of image data while retaining the spatial and image structure. The design of CNNs was inspired by physiology, specifically the local receptive field of regular spatial arrangements of neurons at different spatiotemporal scales. The role of spatially restricted receptive fields for images stems from the fact that most features in an image are not spatially uniform, but sparsely distributed. Furthermore, some features are variations of other features, implying that feature building could be done recursively.

2.1. Architecture of CNNs

The development of Convolutional Neural Networks (CNNs) owes much to the evolution of traditional Artificial Neural Networks (ANNs), which are organized in several layers. The input layer interacts with the input data, while the output layer generates the final prediction. The hidden layers extract features, representing, at each

layer, an increasing level of abstraction [6-8]. Each neuron in the first hidden layer is connected to each neuron in the input layer to model simple features correlated with the applied transfer function, e.g. edges in images. By stacking layers, more complex features that are not easily delineable by a handcrafted process are learned directly from the data. Traditional ANNs are fully connected. That is, each neuron in the l-th layer is connected to each neuron in layer l−1. The number of free parameters is therefore extremely large and grows exponentially with the size of input data. Moreover, in large-sized data, fully-connected layers present the problem of storing high-dimensional tensors. To avoid those problems, in CNNs, adjacent layers are sparsely connected. Two or more layers model a 3D filter that convolves input data. To improve the learning of complex features and to decrease even more the number of free parameters in the ANN, CNNs employ two further strategies.

The convolution operation is paired with a non-linear activation function, e.g. the Rectified Linear Unit (ReLU) layer, that generates a non-linear model. The activation of the intermediate neurons generates the activation map, which is small-sized in comparison to the input data. To keep this tensor small-sized as well in deeper layers while increasing the number of filters that convolve the activation map, pooling layers are used. By using a max-pooling operation, a region of the activation map is compressed into a single number [9,10]. The learned features are local for initial layers and become more and more global in deeper layers. Each neuron of the last CNN layers models a simple linear combination of the feature maps. Classification is obtained through a softmax layer. The basic architecture of CNNs consists of an alternating sequence of convolution and pooling layers, finishing with fully connected layers.

2.2. Applications of CNNs in MRI

Magnetic resonance imaging (MRI) is a non-invasive imaging agent, without the risk of radiation exposure. Through simple volume adjustment of the patient, we can capture a wide range of multi-modal images. In addition to greater anatomical soft tissue contrast than other modalities, the modulation of paramagnetic agents in the MRI scanner can help us detect more detail of the tissue. Based on the absorbing parameters affected by the tissue state, we can capture various image contrast effects including T1 and T2 relaxation time, and susceptibility-weighted imaging. Both qualitative and quantitative analyses can be obtained via MRI. Using deep learning using deep CNNs, which are the most successful applications in the medical domain, we can further enhance the capability of MRI and its application, and we have witnessed a boom in advanced MRI research. To be more specific, CNNs in MRI can be classified as (a) MR protocol planning, (b) Acquisition parameter optimization, (c) Accelerated data acquisition, (d) Data denoising, (e) Motion artifact correction, (f)

Generative modeling, (g) Segmentation, (h) Tissue super-resolution, (i) Disease classification, (j) Multi-modal image synthesis, (k) Quantitative analysis, and (l) Predicting post-treatment consequence [11-13]. However, CNNs are not magical; we need to acknowledge the pros and cons of employing CNNs, and at least one realistic explanation for each study, which should be accessible through the manuscript without additional reading must be provided. In addition, we need to distinguish an academic study on CNNs from other clinical-oriented papers. CNNs have a limitation in terms of research reproducibility and data privacy preservation.

2.3. Applications of CNNs in CT

For applications in CT, many CNNs have been employed for tasks ranging from CT image segmentation of anatomical regions and CT image registration, to CT image reconstruction and super-resolution [2,14-17]. Because CT is a typically low dose imaging modality, many groups have worked on noise reduction of low dose scans using CNNs. Recent works have also explored the challenges of utilizing a limited set of CT projections for the reconstruction of high quality CT images. For tasks such as CT image denoising or super-resolution, the CNNs are trained in a supervised fashion using pairs of low quality CT images as the input and high quality CT images as the output. For tasks such as CT reconstruction from a limited number of projections, however, large scale paired datasets of input-output images can be challenging to acquire. To address this challenge, a physician could make the pairings on a smaller scale and CNNs can then be trained in a semi-supervised fashion, based on synthetic data generated using standard CT reconstruction algorithms. Other variations of CNNs and deep learning generally have been used to augment traditional CT reconstruction techniques through the design of deep priors that leverage the ability of CNNs to learn to exploit statistical redundancies in the data. Such methods, using CNNs, generatively model the reconstruction of the CT image given the retrieving projections as input, as well as the mapping from a high quality CT image to noise or artifacts based on standard CT reconstruction methods as output.

2.4. Applications of CNNs in X-rays

X-ray imaging has diverse applications ranging from security to post-mortem examination. However, the role of X-ray imaging is particularly important in the medical domain. The ease of access and lower cost of X-ray imaging supports its high usage for detecting and diagnosing a number of diseases – dental diseases, bone fractures, pneumonia, and TB. Hence, X-ray imaging presents itself as an attractive target for automatic detection and classification of the underlying diseases.

Deep learning in general, and CNNs specifically, have gained immense popularity in the recent years for the underlying classification and detection tasks in X-ray imaging.

CNNs have been shown to outperform classical methods in a number of detection tasks from X-ray images. In this section, we specifically discuss the application of CNNs in detecting and classifying pneumonia, pulmonary tuberculosis, lung cancer, cardiac diseases, tuberculosis, detecting metallic objects, and in dental radiology. Early implementations using CNNs fine-tuned the weights on the X-ray datasets to detect and classify pulmonary diseases. However, later works attempted to improve upon these results by using deeper CNN architectures. Some of the recent works have proposed either novel CNN architectures, or models pretrained on non-medical datasets, or ensemble of models, or hybrid models, to detect and classify several diseases present in X-ray images.

3. Transformers in Medical Imaging

3.1. Overview of Transformer Models Transformers have gained great attention in not only the domain of Natural Language Processing but also in other tasks, especially in vision tasks [9,18-21]. In this section, we will quickly introduce the motivation of the Transformer model, the original architecture and its applications in the field of vision in general and radiology in particularly. The transformer architecture was originally designed for NLP tasks, which is a model with the main backbone of a sequence-to-sequence architecture that mainly consists of Encoder and Decoder blocks. The training is performed in a supervised manner with the loss functions that are typically cross-entropy loss compared with the ground truth label. The input of the model is a text sequence, which is usually not embedded in the model directly; these tokens are learned to be specific embeddings in the model [22,23]. The input embedding will then go through an Encoder stage, which consists of a stack of N successive layers. Each layer contains two-layer normalization as well as a self-attention block and a fully connected feed-forward neural network. The output of the Encoder stage will then go to a Decoder which is the same as the Encoder stack but with another layer to incorporate information from the Encoder output by performing another attention mechanism. The final output is the target sequence generated during inference via teacher forcing. 3.2. Transformers for Image Segmentation Image segmentation is one of the most widely-studied problems in computer vision [24-26]. The task of image segmentation requires a lot of annotations, thus requiring more attention to build a detailed tool to automatically carry out this task to save time and resources for medical procedure diagnosis and treatment. Various models can be utilized for segmentation methods such as CNNs, RNNs, or more recently transformers. In particular, these models have better speed and accuracy in many common tasks compared to their predecessors. In the world of medical imaging, these methods also deliver comparable results in many scenarios, accelerating the medical process. 3.3. Transformers for Classification Tasks Similar to NLP tasks, vision tasks can also be formulated into text to text format. For classification tasks, such as contrastive learning or more general

tasks such as visual-language tasks, vision encoders are needed to incorporate vision information into the process. The majority of the work utilizes CNNs to encode visual information and then feed the CNN output into a transformer, which was shown to be beneficial for tasks such as visual grounding. For regular vision classification tasks, such as ImageNet, transformer-based models can operate in an end-to-end manner with only a structure encoder only and deliver comparable results or even better results than corresponding CNN models. With the branch of MAE, document clear results can be noticed with masked input to help trained the model in many classification benchmark tasks. 3.4. Transformers for Anomaly Detection Detected such as object detection, change detection, or anomaly detection share similar similarities. In particular, anomaly detection in medical imaging plays a critical role to support doctors to detect and locate suspicious regions in a patient to accelerate or stabilize the treatment procedures if necessary. Similar to document modeling for NLP tasks, sufficient organization downtime is needed. In particular, ViT was proposed to work in a fully vision way, working under the supervision of supervised methods or other testing methods such as zero-shot.

3.1. Overview of Transformer Models

The term "transformer" describes a model topology introduced for natural language processing (NLP) data. In contrast to recurrent or convolutional neural networks, this architecture uses self-attention to relate elements in a sequence. Transformers made breakthroughs in NLP by interpreting texts as sequences of words or word pieces. The original transformer architecture processes sequences with long-range contextualized attention weighing, yielding elegant word representations. These ideas have a techno-cultural resonance that we should explore for object detection in images, volumes, and point clouds.

The original transformer has a symmetric structure, with identical layers in an encoder network that projects sequences into a space of greater dimension in a self-attention-enhanced manner, and a decoder that projects them back into lower-dimensional space for the task of interest: translation, reply, or summary. Each step builds on the result of the previous layer, adding new information [27,28]. In general, the number of layers is much greater than the dimensionality, which is at least double the dimensionality of representations. The layers consist of multi-head self-attention and a feed-forward sublayer. Other differences include Layer Norm protocol (applied at each step before the layer operations, rather than the last step), and sinusoidal functions that could be chosen for position encoding. In images, the information could be delivered sequentially by their sequences of patches rather than pixels. For processing, multiple input sequences are merged into a 2D panorama. Ultimately, what the model "sees" are

the patches. By integrating self-attention with the positional information associated with patches, the transformer assesses patches form a whole.

3.2. Transformers for Image Segmentation

Image segmentation is a dense prediction task, taking a single image as input and classifying the visual content at the pixel level. For certain areas of the image, the class labels may be large, or the area encompasses a large number of pixels. Therefore, the commonly used semantic segmentation loss, such as pixel-wise cross entropy loss, can be class-imbalanced and may overly constrain the network by treating each pixel equally [19,29-31]. Another common approach to multi-class segmentation is panoptic segmentation, which combines concepts from instance segmentation with semantic segmentation. However, densely predicting visual categories can be computationally expensive or inefficient, especially for very high or multi-channel images.

Traditional segmentation architectures, such as U-Net, apply sequential convolutional layers and down-sample predictors to create fused feature representations. Then they unfuse this global representation with up-sampling layers to make dense predictions. U-Net architectures have been historically popular for medical image segmentation, especially when using transfer learning from pre-trained models on the natural image domain. More recently, light-weight U-Net architectures have been proposed, specifically tailored for segmentation of hyperspectral images [32,33]. The medical imaging segmentation literature has seen a sizable number of modifications and variations using U-Net-like architectures. Despite the popularity of CNN U-Net architectures, they do lose information from the input image or predictions at certain resolutions, due to downscaling and up-sampling. Additionally, convolutions, while local, do not explicitly model long-range pixel-pixelized interactions, although local operations can indirectly capture long-range interactions through stacking multiple convolutions.

3.3. Transformers for Classification Tasks

Vision transformers perform classification by first tokenizing an image via non-overlapping flattening short image patches, compute linear projections of those patches which are fed to a standard transformer encoder. An extra trainable embedding for the class token is often added, which is expected to store information that classify the image at the final encoder layer [34-36]. Vision transformers were first properly tested for supervised classification and showed a dramatic increase in performance with scale. Later, this observation was further verified, establishing that vision transformers become universal representation learners with scale. Subsequently, they have also been found to work well in a self-supervised setting via removing the class token, and learning to predict the original pixel values of random patches instead, during training.

The results of the representative work testing the scaling hypothesis for supervised vision transformer are reported, in terms of the error rates for two state-of-the-art classification models based on standard vision transformers and an mlp-mixer that employs dense multi-layer perceptron networks instead of a transformer architecture. It also reports various vision transformer models of different scales, trained on either 300 or 1280 million parameters. When trained at scale, vision transformers achieve significantly lower error rates than convolution-based architectures. Moreover, it also has a very high computational efficiency, in terms of multiply-accumulate operations, compared to prevalent architectures. Vision transformers also learn generalizable representations through supervision, and result in significantly better performance than the standard architectures, when finetuning using linear classifiers after pre-training on large labeled datasets.

3.4. Transformers for Anomaly Detection

Visual Anomaly Detection (VAD) methods can be used to identify and classify unknown image anomalies. One useful tool for anomalous image analysis is a pre-trained image classification Transformer, which has a large amount of image knowledge and can be used to find low-activity regions [37-40]. Previous works extract the image features, identify image prototypes, and extract low-activity image semantic features by clustering. These latent features can then be processed by a memory network. This unsupervised method can find anomalous image detection images.

Compared with traditional anomaly detection algorithms, Transformer works well because of its powerful self-attention image modeling capability and better semantic feature extraction ability from a large image dataset [41-42]. However, at present, Transformers for anomalous image detection methods mainly focus on satellite and forest scene buildings, as well as many naturalistic scenes, but few studies have been done in the medical imagery field. Some apply a transformer visual model to MRI to detect Alzheimer's disease and predict which parts of the structure can represent the activated image. The employ the transformer model to predict Alzheimer's disease with diagnostic short prediction and receding to low activities of MRI. Both of these studies improve prediction by just single-frame classification using sparse vectors at different times and cutting the dataset.

However, few studies have investigated transformer-based memory-less models for the direct prediction of Alzheimer's or other neurological diseases using only static single-frame anatomical MRI data. These classifications, if available, may provide brainstorming diagnostic and prediction tools through preliminary extensive evaluations that identify cognitively healthy people.

4. Generative Models in Radiology

Recent years have seen growing interest in developing generative models of data. Without a clear definition, we think of generative models as functions that take in some latent representation z and return a data point similar to samples observed from the data distribution [6-8]. The most widespread generative model is the Generative Adversarial Network (GAN), which has generators and discriminators competing over a two-player minimax game. The games jointly refine the generator to fool the discriminator into thinking the generated output is a real sample and refine the discriminator to discriminate real inputs from those coming from the generator.

Applications of GANs in MRI are widespread due to the availability of paired datasets and the reported success of GANs in data generation, completion, and improving resolution. Specific cases of GAN use include a study that adapts GANs for rapid MRI with simultaneous calibration. GANs have used built-in regularization methods through the travel distance from latent space into pixel space. This results in the generator building a "latent manifold" mapping low-dimensional latent vectors to corresponding high-dimensional images. Reduction in redundancy helps with bridging the domain gap encountered while training an MR images-to-CT synthesis GAN pipeline.

For computed tomography (CT), GANs have been used to synthesize or denoise images, and also blend different modalities. For positron emission tomography (PET), GANs have been used to achieve truncation artifact reduction in particle imaging. Furthermore, GANs have been used in functional phase detection from untagged cardiac cine magnetic resonance imaging, realistic image synthesis from attributes such as age, gender, and race attribute conditioning. Denoising and improving the resolution of chest X-rays, facial bio age estimation using facial X-ray data, class-conditional photo-acoustic imaging to achieve imaging speedup, data generation against label scarcity, and augmentation of existing datasets are the other applications of GANs in X-rays.

4.1. Introduction to Generative Adversarial Networks (GANs)

Recent advances in deep learning have led to the development of powerful generative models. These models have the ability to learn meaningful data distributions from training data and generate new data that are indistinguishable from the training samples. Unlike probabilistic modeling methods that suffer from one or more of the following shortcomings: suffering from an intractable normalization factor, assuming a specific form for the data distribution, or generating a single data sample from a network, these deep generative models are capable of modeling highly multi-modal and complex distributions, incorporating the learned distribution when sampling.

Among the earliest and more popular deep generative algorithms, Generative Adversarial Networks consist of two components, a generator and a discriminator, trained in a minimax game. Given a set of data samples, the generator maps random input noise from a prior distribution into the data space, while the discriminator is a binary classifier determining whether a sample is from the real data distribution or from the generated data distribution. The generator is trained to minimize the cost function: where the discriminator is trained to maximize the function. This adversarial cost function has a unique global minimum at a distribution that matches the true data distribution. The generator is then used to sample data from the learned distribution.

4.2. Applications of GANs in MRI

MRI can achieve high-quality images with a high contrast-to-noise ratio, especially between soft tissues such as the brain, spinal cord, and connective tissues. However, MRI acquisition is time-consuming due to its physical constraints on scan settings like radiofrequency excitation, selection of pulse sequences, and quantification. A large amount of training data becomes essential for training a deep neural network in order to reduce MRI artifacts. In clinical practice, however, it is typically not feasible to have enough data available for training deep learning networks. Training data are also sometimes unavailable to be used for unsupervised learning. As MRIs are rich in structural information about soft tissues, it was hypothesized that natural images could be leveraged to train generative adversarial networks in an unsupervised manner in order to improve the performance of MRI image synthesis and denoising due to the small number of MRI datasets.

Some work has used generative adversarial networks to develop state-of-the-art solutions for useful tasks in MRI imaging that may overcome these challenges. A few medical imaging tasks that were designed using generative adversarial networks include the following: MRI attribute transfer, MRI to computed tomography synthesis for attenuation correction in emission computed tomography, computed tomography to MRI synthesis, multiparametric MRI generation, MR data augmentation for brain tumor segmentation in radiology, MRI hyperspectral images, high-dimensional imaging, 3D MRI motion correction, and MRI data augmentation for segmentation of pancreatic ductal adenocarcinoma.

Furthermore, generative adversarial networks have developed MRI for a more straightforward reconstruction of low-rank MR images in both 2D and 3D settings while jointly tuning the code space of a method. The method can accomplish MRI in a much less time-consuming manner while simultaneously minimizing any residual noise and without introducing excessive reconstruction artifacts. The MRI generation can also achieve some common motion prior with the help of a generative adversarial

network, which provides data-specific supervision and alleviates the large pose-gap problem encountered when training conventional motion correction networks.

4.3. Applications of GANs in CT

In CT, GANs have displayed a remarkable capacity for sophisticated image synthesis. Their accomplishments encompass a multitude of tasks, including consensus image reconstruction enhancement, instrument deficiency correction, denoising, error diffusion rectification, and enhancement of superior resolution through super-resolution. Furthermore, GANs have also garnered attention for their unique capability in image-to-image translation, which enables the mapping of sketches to CTs and the transformation of CTs into photorealistic counterparts. While a significant portion of GAN applications in CT has been predominantly experimental, their plausible effectiveness has sparked interest from both the industry and academia in translational GAN research. Such intrigue bodes well for the promise of GANs throughout radiology.

Coherent image reconstruction from limited views has long been a focus of research spanning over two decades. A proposal was illustrated to remedy the drawbacks of a conventional model and synthesized a better image from limited views utilizing a GAN-based strategy. Conventional images are reconstructed using a clinical model, and the artifact component is eliminated through a zero-padded limited technique. Subsequently, the images were synthesized by training a GAN model. Finally, additional images are generated with high artifacts that belong to previous years. Afterward, a GAN model is trained based on additional images. Once the control GAN model is trained with pre-augmented loss, it is applied to novel images, thus decreasing artifacts on limited images and synthesizing coherent images from limited views.

4.4. Applications of GANs in X-rays

Although X-ray radiography is the earliest invented medical imaging modality, it is still widely used in clinical practice. Also, its important medical applications, including medical diagnosis, human skeleton identification, and plant disease, meet the challenge of low-quality and insufficient X-ray training images because the X-ray imaging process is sensitive to equipment and environments, which increases the risk of X-ray misdiagnosis. It is known that supervised learning based on convolutional neural networks requires a lot of labeled training data and performs poorly when the amount of training data is insufficient. Generative Adversarial Networks can generate new data and enrich labeled data to mitigate these problems. Therefore, GANs are widely applied in diverse X-ray tasks.

In the medical domain, GANs can be used for generating high-quality X-rays; generating labeled unpaired X-ray images; converting other modalities to pseudo-X-ray data; data denoising; and data augmentation. As for the commercial domain, GANs can be applied for improving the quality of low-cost X-ray imaging; improving the performance for the task of abnormal class detection in vertebra X-ray images using unsupervised domain adaptation; and synthesizing specimen X-ray images to develop an automatic explosive detection system. These very diverse X-ray applications verify the wide application of GANs in commercials and clinics. The advent of GANs provides us a new way to tackle difficult X-ray tasks, and it will further improve the performance of those fields, such as the X-ray classification, segmentation, and reconstruction tasks.

5. Image Segmentation Techniques

Image segmentation classifies each pixel in the image according to some predefined labels. Often in natural images those labels are related to objects present in the image. The result of such procedures are structured representations of complex visual data where low-level features such as color or texture are grouped. Segmentation is often one of the first stages of visual perception. Perceptual studies indicate that the human visual system first extracts the edges of objects and areas of uniform characteristics. The initial image is first divided into large areas and then refined. In segmentation applications in medical imaging, there exist similarities to those found in natural images. For instance, structures of interest have often low contrast, and similar features also exist in areas which do not represent structures of interest.

In medical imaging, pixel-level labeling is crucial. For instance, in CT and MRI, segmentation largely determines the correctness of diagnosis done by radiologists, and thus a number of algorithms have been proposed focused on that task. For instance, in CT applications, shape features have been widely used to create procedures that are invariant to low-contrast appearance. Others have proposed using mathematical morphology or dynamic programming using the 2D surface representation, which integrates local features with 3D shape priors. In general, however, such methods are limited and supervised pixel labeling has become a must in many applications. Deep learning has emerged as one answer to those limitations where the learning algorithms are trained in manually labeled data sets. The early days of convolutional neural networks were focused mostly in categorizing patches extracted from the images.

Requirements in the medical imaging field have moved early work in the general fields of semantic and object category segmentation to their instance level segmentation. For instance, recovering individual bronchial trees from 3D chest CT is very important for diagnostic and therapeutic tools. Therefore a set of algorithms that segment individual

objects have widely been applied in that field. These instances refer to individual patient-specific anatomical components relevant to some task, so that instance awareness is important.

5.1. Semantic Segmentation

Pixel-level pairwise labeling of images is probably the most elementary level of understanding a visual scene, and humans do not think twice to assign pixels into pre-defined classes and to carve-out free-space holes around building faces, assign pixels on trees' leafs with a yellow-colored class label associated with autumn, or even risk assigning labels into shadowed zones when doormats create a spatial connection between indoors and outdoors. Nevertheless, such elementary labeling tasks are challenging ill-posed image understanding functions previously explored in the context of hierarchical image features and post-processed hand-tuned classifiers. Medical image segmentation aims to partition images into different structures associated with relevant anatomical structures or scanners being used and pathologies being scanned. Note that most types of medical scans are inherently 3D or even 4D, but in the imaging community it is common to slice or flatten space for a surface-volume model into a 2D image, and this is reflected in the standard nomenclature when referring to the involved scanners. For MRI modalities, the structures usually of interest are anatomical structures such as gray or white matter associated with the outer layer of the brain or even deep structures such as thalamus or basal ganglia; ventricle cerebrospinal fluid spaces; and skull and facial bone, vessels, scalp, and cartilage regions. For CT, the anatomical structures of interest are almost always the same as in MRI apart from the CSF region; in addition, bone pathology and, to a lesser extent, cartilage pathology. These structures are typically imaged in problems such as segmentation atlas computation or enhancement.

5.2. Instance Segmentation

Instance Segmentation (IS) has become a trendy theme in segmentation research, it jointly predicts object category and delineates object instance mask. Most of the methods follow the two-stage pipelines, extending Faster R-CNN by adding a low-resolution segmentation branch. Randomized labels on segmentation RoI feeding, Input Feature of nonlinear mask prediction layers, and adaptive reshape of the mask losses form a mask-loss boosting strategy for instance segmentation. A box-Segmentation Refinement Network (bSRN) simultaneously refines the class-agnostic box input and the class-conditional segmentation output at each stage, guiding segmentation refinement tasks toward faster convergence by prolonging the intermediate segmentation, and expediting both inference and training via a multi-stage

framework. Considering IS does not need fully supervised support, a self-supervised depth-driven instance segmentation framework for monocular scene was explored.

In multi-modality applications, a spatial information-enhanced multi-modality one-stage instance segmentation method for multi-GBSCT data was proposed. However, few researchers attempt to tackle instance segmentation in medical imaging domains, while an end-to-end designed instance segmentation architecture for discovering and delineating different anatomical vessels and branches from 3D fundus images of different retinal orientations was proposed, detecting anatomical blood vessels and masks leveraging Sn-Att U-Net on fundus images. Mask Skin U-Net was adopted for both automatic segmentation of facial skin lesions in dermoscopic images and accurate identification of background, increasing the predictive accuracy of the model in IS task. It was proposed that Mask R-CNN was utilized to detect and segment fetus from the obstetrical ultrasound images. Given the unique challenges associated with the sensitive and subtle changes in the anatomical structures at early stage of cardiac in prenatal ultrasound, an instance segmentation based shape-aware deep-learning algorithm for semi-automatic segmentation of fetal cardiac limbic was introduced, achieving satisfactory accuracy and generalizability.

5.3. Comparison of Segmentation Approaches

In this section, we compare the merits and demerits of the segmentation approaches discussed in Sections 5.1 and 5.2. We begin our discussion with a review of semantic segmentation approaches. Semantic segmentation is performed using multi-class supervised techniques. The segmentation maps of all classes are initialized, i.e., the pixel values of the segmentation maps corresponding to all classes except the class of interest are set to 0. The model is trained to minimize reconstruction errors for only the given class, while not affecting the pixel values of the segmentation maps corresponding to other classes. It is important to note that, in case of multiple instances of the given class, the shaded pixels must share the same class. Regression-based object detection models require that the pixels of a dense segmentation map for the current image and the 2D Gaussian maps of all objects in the image share the same class. The pixel-wise response must be a single density value corresponding to the class of interest because, on a 2D Gaussian map, higher density values signify that the object is closer to the camera. Also, when the distance is near the camera, the object is larger in the image than when the distance is far from the camera. The semantic segmentation tensor computed by a model trained using multi-class supervised multi-instance segmentation dataset has density values that satisfy the above property.

Unlike semantic segmentation, instance segmentation is performed using supervised one-class segmentation techniques. Unlike dense detection approaches, instance

segmentation approaches segment instances of only the given class of interest. The trained instance segmentation models minimize reconstruction errors for only the specified class while leaving the segmentation maps corresponding to the other classes and pixels, which are location-stationary, unchanged. Therefore, the trained model does not overwrite the computed segmentation details of other classes. After instance predictions are computed on the input image, the additional object counts, segmentation maps, and other image cues such as occlusion, vignetting, and lighting can be performed on the instance predictions to promote more accurate multi-instance segmentation.

6. Image Classification Techniques

Deep learning techniques are widely applied for medical image analysis problems of various nature. The techniques allow the analysis of radiological data concerning multiple anatomical parts and a variety of different diseases. Often, medical imaging techniques allow the collection of rich and detailed information about the body of a patient and current deep learning technology is aiding the fast and efficient classification and analysis of the collected data by replacing more traditional approaches utilizing rule-based systems. In this chapter, we explore the use of convolutional networks, a type of deep learning architecture that has been successfully applied to various analysis problems in both natural and artificial medical images, outlining the current state of the art and the possible short-term foreseeable future.

The image classification task is predominantly framed as a supervised learning problem. The objective is to learn a classifier which, given an image in the input space, a probability distribution over labels in the space, and a loss function, optimizes the network parameters, or weights, so that it minimizes the average classification error across the entire training data. The trained network can then be applied to new images which it has never seen by performing inference with the trained parameters. It will output a predicted class label for each test image. Supervised image classification approaches can be mainly subdivided into different approaches, depending on the number of classes defined in the task, which can be binary or multi-class.

6.1. Binary Classification

Binary classification represents the most fundamental and simplest image classification task to perform, where a classifier distinguishes between two classes of images, given the training images from both classes. This premise also forms the foundation of all other image classification approaches, including multi-class classification, where the classifier distinguishes between multiple classes or their parts of images. Multi-class classification is implemented through the use of the binary classifiers that independently distinguish between images of one class and the rest of the classes or

combined into a single optimization function that implements the required one-vs-rest or directed acyclic graph concept. Many image classification problems require the ability to identify only the images of a specific class. Image retrieval attempts to find those images in a large database determined and categorized by a set of criteria.

Binary classification is one of the tasks that supervised learning techniques such as deep learning can solve. At its core, a supervised image classification task has both image and label classes, and using the information from many labeled examples, is able to assign to new unlabeled examples a label class. While supervised image classifiers can take many forms, including heuristics, statistics, and mathematics, the most successful image classifiers in recent years have been based on deep learning convolutional neural networks. In its simplest form, CNNs take as input an image and output an anticipated class label. Given training data in the form of labeled examples, CNNs learn the correlations between the input image data and the output class. Once trained, CNNs can classify new unlabeled images in the same manner, allowing us to categorize or segment images.

6.2. Multi-class Classification

Multi-class classification aims to assign to an input image the label corresponding to one among a number C of possible classes. In its most fundamental form, classification is a per-image task, meaning that a classifier network outputs a prediction relating to the whole image. For such a per-image classification setting, one usually uses a classifier based on feed-forward networks to extract high-level features from the image, which are then fed to a multi-class softmax layer predicting the probabilities of each of the C possible output classes. A classifier's final loss is based on the following categorical cross-entropy loss:

Inherent in using the categorical cross-entropy loss is the assumption that image classification is an independent task where the model has to predict the class probabilities for the whole image alone. This is generally the case for image-level tasks such as disease classification, where the model determines whether a specific disease is present in the chest radiograph. Some diseases can be assigned a class based solely on information in the image but that does not necessarily require the model to consider the whole image context. However, when classifying histopathological images, an image is usually scanned tile-by-tile. Each tile is classified based on visual similarity against tile images annotated with different class labels of tumor types, tumor composition, cancer region quality, or tissue type. In this case, the label of the complete image is obtained based on the class label that was associated with most tiles.

6.3. Performance Metrics for Classification

Classification is one of the most prevalent tasks in image analysis. It typically involves training a classifier on a collection of labeled images, and then utilizing it to assign the label of each image in a disjoint set of test images. Given the scale of modern datasets, with several hundreds of thousands or millions of images, it is common for classifiers to achieve errors in the 10−15% range. The task is considered solved for certain datasets when the error achieved by a classifier is better than the error achieved by an average human annotator. The vast domain of supervised image classification has also been addressed with far fewer training samples not far from its feasible limits for certain applications. Not only does the average performance of classifiers on classification tasks improve as systems scale but the range of tasks to which supervision can be applied has also become diverse: from distinguishing individual objects to identifying fine-grained variations within species categories. To a lesser extent, the domain has also expanded through few-shot, and even zero-shot learning, where classifiers are trained with as few as one labeled image. Within the medical imaging analysis community, classifiers have been utilized to perform tasks ranging from whole image classification to pixel-level diagnosis.

A major distinction between the supervised image classification tasks in the wider computer vision community and those in medical imaging is that the former are balanced both at label and dataset level: All labels comprise roughly the same number of images and datasets are sampled from the same joint distribution. Such a setup is ideal in that all classifiers receive the same amount of label feedback. In contrast, medical imaging classification tasks are often highly imbalanced. An example is x-ray image classification, where several x-rays of pathologies would be labeled with the same abnormality class, yet many healthy x-rays would belong to the non-pneumonia category. Such setups pose both practical and philosophical difficulties: First, the data imbalance sometimes results in classifiers being ineffective at the abnormality classes, thereby undermining the usefulness of the system. Second, the classifiers cannot be interpreted as deriving from a good-feedback learning principle. In these cases, classifiers trained with expert supervision are more akin to mimicking the collective behavior of the experts than actually deriving from their judgment.

7. Anomaly Detection in Medical Imaging

In this chapter, we review recent advances of anomaly detection methods, including classical, GANs, and self-supervised-learning based approaches, using different types of images ranging from X-rays to 3D MR scans. A common objective of medical imaging is to support diagnosis of a disease or the state of a patient. Structured lesions or a certain connectivity of directly or remotely related tissues is usually sought in

different imaging modalities in order to anchor the diagnosis. However, most often, anomalies do not appear as local lesions and are not confined to nearby regions in images, but share a complex spatio-temporal structure with healthy tissues and may involve functions other than image intensity or color. The structural relationship with healthy tissue can be used to reject other pathologies or abnormal functions in static images and videos, and can assist the doctor in assigning a disease to an abnormal state without a well-formed hypothesis about the features to track.

These properties suggest that for unconventional diagnostic cases where few images or established diagnoses are available, unsupervised anomaly detection schemes can be applied. Indeed, while some of the popular supervised or weakly supervised approaches for common lesions perform well, the cost of annotating and training the models for the less frequent diseases may exceed the benefit, as the models may just memorize the lessons or recurrent failure cases. Instead, methods for conditional generation of images, reconstructive CNNs, or even generic SSL approaches have been trained without paired or partially labelled image datasets, but rely instead on rich redundancies of the underlying anatomy and/or pathologies in the analyzed dataset and may collapse given badly calibrated models. These paradigms have found increased prominence in medical imaging practice, after years of success in on-line services in face, object, and scene recognition, effected by the rapid development of computing power, the availability of new large-scale unlabeled image and video repositories, and the self-organization of image categories by consumer and research users, among others.

7.1. Techniques for Anomaly Detection

Specific techniques for anomaly detection can be divided into three categories: unsupervised learning, partially supervised learning, and supervised learning. Three primary deep learning techniques are commonly used in medical imaging, which also led to many computer-aided diagnosis engines applied in clinical practices. The first category is unsupervised anomaly detection, which uses the normal data distribution to identify an anomaly image as an input. The simplest method to implement anomaly detection in an unsupervised manner is to use traditional image filters. A simpler form would be thresholding, which assigns colors to pixels based on band ratios or filters using a convolutional operation. Expert-designed filter methods can be quite successful if sufficient knowledge of anomalies is present when formulating the filtering rule. However, many anomalies are subtle differences in pixel values in a background of normal samples and might therefore be unnoticed. Thus, recent research efforts have looked into learning a model to distinguish between normal and anomalous images. Such an approach requires a training dataset, distinguishable by the computer or a model parameterized by human experts. For instance, the autoencoder is a neural

network trained on a set of samples of only normal data. The second group, semi-supervised anomaly detection, uses both normal and abnormal samples for model training, such as supervised textures. This technique can gain greater discriminability for the classifier due to the supervised nature of model training. The most superior yet demanding way is supervised learning, which uses plenty of anomaly samples for training and generally produces the best model since it uses the most amount of information in both normal and anomalous subject datasets. Such methods have the additional downside of requiring many annotated samples and, therefore, challenge the practicality of supervised learning, especially in the medical imaging field, where the number of patients with certain diseases is meager compared to the normal patient count.

7.2. Evaluation of Anomaly Detection Models

Anomaly detection in medical imaging has a long-standing history in research and application. Anomaly detection has more recently been revived by the significant increase in curated visual data enabled by recent advancements in deep learning. Reassuringly, the importance of establishing suitable evaluation protocols for anomaly detection has also been acknowledged by deep learning researchers. Reviewing a diversity of evaluation strategies, we summarize the various ways in which anomaly detection models are evaluated and how these connect back to the fundamental task and its underlying goal. Through this process, we will notice similarities as well as differences between existing evaluation protocols. We categorize these evaluation strategies based on the amount of data and the type of evaluation heuristics used. The first group requires more than one dataset, while the second one relies on a single dataset. The second group is usually based on supervised metrics, such as classification scores. Moreover, our examination complements previous evaluation efforts by providing extensive coverage of evaluation methods. Notably, most recently proposed adversarial or self-supervised models do not use an external labeled dataset for evaluation. Hence, we explore several of such techniques to evaluate various types of anomaly detection methods without a predefined labeled dataset.

We first categorize available medical images based on modalities, imaging process, dimensionality, and presence of anomalies. Then, we explain all the possible requirements and wishes of the user performing the evaluations and present a table that summarizes the whole evaluation process and its aim. Our literature review and table address the four main user wishes; fastest evaluation, most available data, most data with anomalies, and most resources to perform the evaluation. Many research works present similarities regarding the modeling and evaluation of the pipeline. Although these pipelines are addressed by either using specific phases, model names, or summary metrics, they share similar notions. We also summarize in a table all

available techniques and show the main phases that are either skipped, focused on, or summarized.

8. Data Augmentation Strategies

Deep learning methods require a large number of datasets to avoid overfitting when training models for specific tasks. Most tasks in the medical domain only have a limited amount of labeled imaging data. Data augmentation can artificially expand the size of training and validation datasets to a large extent and so, has become the demarcating factor between models that perform well on the training set and those that generalize across testing datasets as well. Various augmentation techniques such as rotation, flip, shift, intensity transposition, and elastic deformation have been shown to improve the performance of medical imaging tasks such as detection, localization, segmentation, and classification. In contrast to such CNNs that use generic image datasets as a pre-training stage and medical datasets as the fine-tuning stage, the initial layers of CNNs trained on larger scale and task-specific medical datasets learn filters that are specific to the training data.

Boolean operations such as AND/OR are important for generating binary datasets required for a specific segmentation task. However, the above-mentioned techniques do not ensure the validity of the segmentation labels. GANs are a possible solution for generating datasets with valid segmentation masks. In summary, data augmentation improves the model's ability to predict imbalanced datasets and makes it less sensitive to noise. It also encourages the model to explore more unlabelled regions of the dataset, and a combination of different augmentation techniques produces the best results in most use-cases.

8.1. Techniques for Data Augmentation

Data augmentation techniques fall into two general categories: image based or data based. Traditional augmentation methods such as rotation and flipping manipulate images; these fall easily under the image based category. By contrast, more advanced data based methods model images directly in the dataset space by generating a new simulation encompassing image data from a model that describes image generation. Each of these categories has advantages and disadvantages; most use cases appear to be image augmented, for several reasons. From a technical point of view, image based data augmentation is often easier to implement and faster; it further accounts for widely used simulated datasets. Clinically, the computational simplicity of image based augmentation can also be an advantage: clinical workflows rely on fast image transfer and use; image based data modifications are easier to implement in this setting than complicated data space modifications. Using synthetic datasets for training is therefore an appealing way to reduce sensitivity to domain shift, but generating large

scale datasets for a specific clinical context requires time in the work up and the cost of high performing computer systems.

The most basic methods, including shifting, cropping, mirroring, flipping, rotation, elastic deformation, and adding noise, can only generate new images sparsely related to the original data. Advanced techniques, such as generative adversarial networks and the Bilinear Generative Models implementation, or diffusion based data augmentation methods are able to generate data which is visually so close to the original ones that observers can't distinguish them. Yet, years after the first successes in GAN-design and application reporting, introducing data volume augmentation through GANs in clinical practice is an open question; GAN-methods suffer from still relatively high mode collapse characteristics which thus needs to be countered with further post-processing of the augmented datasets.

8.2. Impact of Data Augmentation on Model Performance

Identifying the best Data Augmentation configuration is not easy. Ideally, we would have a very large and diverse dataset including all possible variations characteristic of the underlying population, The main advantage of DA using synthetic transformations is that we can artificially generate as many examples as required, thereby bridging the data distribution gap. As the model starts training on this larger dataset, the accuracy increases with the number of images that utilize the new magnitudes added to the loss function, at the expense of a so-called double descent Risk curve. In this chapter, we discuss some of the best practices for DA for improving model performance while mitigating overfitting. Some of these ideas stem from practical experience and others from theoretical arguments.

Several major attractive ideas can be highlighted from empirical evidence regarding DA techniques and model training. First, most augmentation strategies are particularly valuable for specific image types, for instance, photometric modifications are often useful for natural or "everyday life" images, rotational paradigms are to be preferred for objects that exist in a three-dimensional space, etc. Second, different DA methods can be combined to good effect, e.g., horizontally flipping an image from a color distorted augmented pair of images can further help the model learn. Finally, there are different levels of DA techniques; some methods apply while the model is optimizing, while others apply at a higher level and can thus be twinned with other augmentation methods.

9. Handling Data Imbalance

Data imbalance is a significant problem associated with training machine learning models to classify medical images. Although machine learning models generally

become better the more data they are provided with, they can fail to learn at scale if the proportion of samples in the minority class is too small. For example, there may be thousands of images depicting healthy patients, but just a few dozen images of patients with a rare but dangerous illness, with more examples teaching the model what a normal image looks like while very few images show the normality to vary on. The model might respond by classifying all unseen test images as healthy in order to minimize its classification error rate, which is a strategy that fails singularly in respect of the minority class, where any misclassification should blacken the model's record.

While general machine learning considerations call for growing the number of training samples, growing the number of images for only the minority class in medical imaging does not have a practical solution most of the time. Underlying this reality are the costs and risks connected to the complexities of acquiring image data. Deep learning also exacerbates the problem of data imbalance because of its need for vast amounts of data for probability density estimation. Moreover, the associated predictions are often described in terms of the quantities required for model training. It is interesting to note that the accessories used to take images of rare diseases may themselves classify as minority classes. For instance, images of rare bone disorders in dental radiographs may not only represent anomalous medical conditions but may also be characterized by the presence of specialized equipment at identical test locations.

9.1. Techniques for Addressing Imbalanced Datasets

Some of the many applied strategies to deal with imbalanced medical datasets include stratification, augmentation, distinguishable performance metrics, reweighting/cost-sensitive learning, targeted correctioning, hybrid classification, and oversampling methods. In addition to these special methods, one could also apply traditional solutions: data collection, transfer learning, ensemble learning, semi-supervised learning, active learning, and the system design correctioning.

Stratification is often the first way of handling imbalance before model training by preserving the proportion of each class during splitting. Since deep learning models require large datasets, it is often infeasible to increase the data, particularly for the minority class, which brings the necessity of augmentation. Targeting the underrepresented samples in the pre-processing can be proven useful, which can be performed using different methods dependent on the model and the nature of the data. Examples include using different augmentative transformations for the different splits, using lightweight augmentation for easy samples, or applying parametric extrapolation only to the low-density regions. The use of distinguishable performance metrics in the training process can be proven useful but requires good care in the selection of these metrics. It is worth mentioning the pitfalls of precision and recall, for example, due to

relying on the low-overlap region. A proposed reweighting scheme modifies both the probability losses and the spatial softmax function used in multi-label segmentation.

Deep learning uses large amounts of data, thereby rendering sampling misrepresentation correctioning infeasible. Deep learning also uses multiple stacked layers of low-level descriptors, which causes the system to have increased robustness against label noise. Due to these advantages of neural networks, traditional solutions adapted to a transfer learning based system architecture and model can be proven useful. These conditions mainly detail data collection, data merging, ensembles, semi-supervised learning, and active learning.

9.2. Effects of Imbalance on Model Training

The problem of data imbalance at the training stage makes it difficult to use common ML for MI since model performance will be too sensitive to incorrect labels in the minority label types. First, the usual objective functions minimize the differences between predicted probability distributions and label probability distributions, which are based on entropy and modeled expected test performance under the assumption that both label probabilities and input distributions in training and test are the same. This assumption is often not satisfied in practice for MI. Therefore, the small labels' probabilities and the large input distribution corresponding to the small label probabilities can trivially affect the performance of the proposed objective functions for highly imbalanced datasets.

Second, the standard loss functions inducing softmax score functions estimate the probability distribution of the model input based on a large amount of training data. Since the input distribution based on model training as generally observed for the large probabilities will be dominant at the beginning of model training, model weight update at this stage would be too sensitive to incorrect labels in the small label classes. The large softmax score function corresponding to the small label probabilities indicates small model input likelihoods, which will contribute a trivial amount to model weight updates since they would not spread model behavior to the small label classes. This effect may also extend to fine-tuning stages. In a nutshell, if we do not adapt or modify the conventional ML methodologies for MI to the problem of data imbalance, these formulations and considerations show that the performance of the proposed algorithms may not be very good on MI tasks with highly imbalanced datasets. Adaptively augmenting the MI label space adaptively without label space expansion would help overcome this problem.

10. Generalization in Deep Learning Models

Generalization measures the quality of the learned mapping between the input and output spaces, when the input data are drawn from a distribution different from that of the training data. Generalization is one of the most important aspects of learning as our ultimate goal is to perform well on unseen data. Generalization quality is more important than the training performance, as even a poor-fitting learned mapping can give good predictions for unseen data. Poor generalization may be due either to overfitting or underfitting. The learned prediction can be too complex and approach the Dirac delta function centered on the observed outputs associated with the training inputs. In this case, we say that the mapping is overfitting the training data. Overfitting can occur even when the training error is small due to a model with many adjustable parameters and strong flexibility. Conversely, the learned mapping can be too simple. It misses the needed structure in the data and results in similar predicted outputs for input data from the training set and from outside the training set. In this case, we say that the learned mapping is underfitting.

Generalization applies to all types of learning, not only supervised learning. The functioning of all learning mechanisms is strongly affected by generalization. All practical systems allow an essential variation on the inputs during their development; they thus require capabilities, built-in or learned, to correct effectively the deviations from the expected results. We can take advantage of the available data and of the learning process by designing it in a way that prevents the construction of too flexible solutions that span the whole possible input range. Generalization in supervised learning algorithms is especially important. From a predictive and practical point of view, the ultimate goal of algorithms that employ some form of learning during their operation is to be successful in producing the required output, given a new input value that has not been presented previously.

10.1. Overfitting and Underfitting

Despite their remarkable result, deep learning models still face the generalization problem. Generalization describes the ability of a model to make accurate predictions on unseen data. In the context of supervised learning, given a training dataset comprising labeled examples, the model aims to learn the underlying mapping rules that relate the input samples to their corresponding labels. The generalization error measures the difference in performance between the training data and the test data that contains new samples not present during training. A model that would achieve a small generalization error would be considered as one that has good generalization capabilities.

In contrast to classical training strategies, which minimize the training loss while measuring the generalization capabilities using validation datasets, deep learning models proficiently utilize the large training datasets to minimize the training loss with little consideration for the generalization. This lack of concern for generalization can lead to two extreme situations in which models are unable to accurately predict labels for samples in the test dataset. The first situation, known as underfitting, occurs when a model does not learn the mapping rules even after prolonged exposure to the training data. This is more likely to happen when the model has a simple architecture, commonly defined by few layers or filters. In this case, both the training and test error are large. The second situation is called overfitting. It occurs when a model becomes too efficient at learning to accurately predict labels for the training data with very low training error, but becomes ineffective at generalization, resulting in a large test error.

10.2. Techniques to Improve Generalization

One of the simplest ways to prevent overfitting and improve generalization is to augment the training dataset by modifying the training samples. A few standard augmentation techniques include randomly cropping, mirroring, and rotating images, adding random noise, color distortion or blurring. Such techniques are especially popular in image classification and detection tasks, and their utilization is almost mandatory in problems involving small datasets. Further, it has been shown that specific augmentations, if applied during training, also improve the robustness of the model. Augmentations also increase training time and model training with augmentation is difficult to converge.

Dropout is a visualization tool that helps minimize the overfitting of DNNs. In particular, it has been found to be very effective for visualizing deep networks at the fully connected layer. The dropout method creates a randomly sampled subset of hidden units to probe the network. The method seems to add subsampling noise to the learning, which is similar to model averaging, leading to a better generalization performance. Additionally, skipping a uniformly sampled set of hidden neurons mitigates the overfitting by preventing the interpolating behavior of a deep network. At the test time, the output becomes the average of a set of dropout samples from all possible samples. Uncertainty introduced by dropout encourages the active learning effect, and in practical deep learning models, this approach is used whenever the model is limited by the training data.

Indeed, the original goal of dropout was to mimic multiple processes during both training and testing of a single small network. However, a naive use of dropout in practice can backfire, with current models requiring learning rates half as small without dropout. Thus, dropout might not be declared the way to go for all visual

manual models, even though it generally leads to better visual results than a non-dropout model. Despite this, dropout remains a popular mainstay among visual models.

11. Future Directions in Deep Learning for Medical Imaging

The advancement of medical imaging provides the physician with high-quality, clinically relevant, and timely information, which helps in both diagnosis and assessment of disease burden, and treatment planning and response assessment. We envision further rapid development in the medical imaging field with increased data sharing, data standardization, better and more ubiquitous compute infrastructure, better and more advanced feats of curated datasets, unified and open-source model libraries, and creation of a community across the applied and clinical scientific space. Better access to data provides a fertile ground for data-hungry modern deep learning architectures. Procurements of more powerful, lower cost, and efficient imaging devices will make acquisition easier and widespread and the easy availability of such data may help narrow the gap of generalization of models to non-institutional datasets. Better and reliable tools for data annotation will help with generating curated datasets with ground truth labels needed for supervised training, pipelines for semi-supervised and unsupervised training will help alleviate the need for labeled data, and efficient validation of such methods will allow for testing domains like self-supervised tasks to generalize to various downstream supervised tasks. Increased partnering between industry and academic institutions and data-focused consortia will help in the creation of well-curated repositories of focused imaging datasets, readily allowing for method research and prototyping to become part of the training of the next-gen innovative workforce. Environments that allow for rapid model development and prototyping, such as easy-to-use model libraries, and open-source implementation sharing, will foster creativity and expansion of deep learning application development to the medical imaging field. Part of the next wave of innovation should include bridging the gap between imaging data analysis and clinical needs.

12. Ethical Considerations in Medical Imaging AI

Concerns have been raised regarding bias in automated decision-making systems. Risk of biased performance can arise when such systems are trained on non-representative data. Moreover, unequal performance can also pose a risk with medical imaging AI, even when the distribution of training data is well-balanced and is representative of the real population. Bias in performance can have additional consequences, including the potentially biased allocation of medical resources, cost-saving, equity and human rights issues, reductions in public trust in medical systems, and constrained algorithm performance when finetuning on small datasets or when used in different regions, hospitals, or clinical workflows. Public health discourse regarding equitable patient

outcomes is therefore relevant in the context of medical imaging AI. To put the AI algorithms into service, it is crucial to make them equitable. Equitable performance across patient demographics is only achieved if considerations, testing, and validation are built into the algorithmic design from conception onward.

Medical imaging AI may also change clinical workflow in such a way that its use disrupts existing frameworks for the informed consent of patients or alters the doctor-patient relationship. AI methods should produce explanations.

13. Conclusion

Medical imaging is regarded as the eye of modern medicine. While medicine has uncovered a huge knowledge challenge, this cannot be automated due to the absence of automation in today's medical research or suggestive intelligence from technical advancements in Artificial Intelligence. Simultaneously, medical research has accumulated a huge amount of medical literature; however these articles are often not linked or no contextual correlation is made. It is a big challenge for practitioners to dig into the entire available medical scientific literature to provide the appropriate medical help in certain cases. Digital choices can be further supported by the quest for knowledgeable and AI-based options for diagnosis assistance and therapy programs. The different forms of accessible medical data can be perceived as a treatment of the patient which is substitutive to the medical work and usually ethics seek the consolation in the bond between the doctor and the patient. Despite all progressions, innovations or enhancements, a doctor is not disposable. Yet, AI has significantly enhanced productivity and services in recent decades.

Deep learning transforms the practice of computer vision. The newest methods put together large general differentiable neural nets with extraordinary levels of optical characteristic and then get to learn the thousands of parameters in these networks, typically using enormous modern data sets. Notably, these approaches are capable of automatically discovering interesting patterns in data without requiring a human for engineering difficult features which have characterized prior generations of pattern recognition methods. This has had a remarkable impact on the recognition of faces, objects, scenes. What has not so far been successful with these types of approach is the recognition and analysis of the patterns of activity that unfold in time and space. In considering this issue, we must note that seamlessly ongoing experiences from stereo spatial locations allow not only 2D visual sense, but also a 3D visual sense coupled with reactions from an additional "sense" - subseasonal statistics.

References

[1] Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. CAAI Transactions on Intelligence Technology. 2025 Feb;10(1):1-35.
[2] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.
[3] Khan SU, Asif S, Zhao M, Zou W, Li Y, Li X. Optimized deep learning model for comprehensive medical image analysis across multiple modalities. Neurocomputing. 2025 Feb 28;619:129182.
[4] Santini F, Wasserthal J, Agosti A, Deligianni X, Keene KR, Kan HE, Sommer S, Wang F, Weidensteiner C, Manco G, Paoletti M. Deep Anatomical Federated Network (Dafne): An Open Client-Server Framework for Continuous, Collaborative Improvement of Deep Learning–based Medical Image Segmentation. Radiology: Artificial Intelligence. 2025 Apr 16;7(3):e240097.
[5] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.
[6] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.
[7] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.
[8] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.
[9] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.
[10] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.
[11] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.
[12] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.
[13] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.
[14] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.
[15] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.
[16] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.

[17] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).
[18] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.
[19] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[20] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[21] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[22] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[23] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[24] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[25] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[26] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.
[27] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.
[28] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.
[29] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.
[30] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.
[31] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.
[32] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[33] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.

[34] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.
[35] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[36] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[37] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[38] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[39] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[40] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.
[41] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.
[42] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 5: Interpretability in Clinical Artificial Intelligence Systems

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Interpretability in AI

Artificial Intelligence (AI) is a broad field that engages many fundamental questions about learning, cognition, inference, and intelligence itself. Utilizing techniques such as neural networks, AI has achieved great successes in trusted problem domains like facial recognition, image classification, language translation, and health diagnostics. However, these systems are rarely ever applied without some reservation. One of the foremost weights dragging down the optimism surrounding AI is the lack of uncertainty quantification; when the world diverges too much from the training data, AI cannot clearly signal its impotence [1,2]. That said, AI can still misclassify in quite specific, explainable ways, seeming to reflect its own logic. For many AI applications, this level of learning is not good enough; we require an understanding of how the system arrived at its answer. The desire to make practical use of these unreliable systems has led to increasing interest in interpretability; we want to be able to look at an AI's computations and use that information to help understand how the system can provide us better information. In this essay, we explore interpretability in the narrow and broad sense, in images and in dependencies [3-5]. For some AI applications, a fidelity-based interpretation may not be sufficient. Instead, we are interested in understanding AI's solution pathway; we want to illuminate the reasoning of a decision made by the AI, even if that logic differs from our own. Such procedural understanding appears to be a step to scaffold a user's decision-making process or to solve a problem too difficult for a lone human or AI to solve. We're curious which AI frameworks are positioned best to address the interpretability deeper challenges of solution pathway illumination and shared decision-making scaffolding. We survey categorizations of

interpretability definitions and methods, with the distortion functional formalism to provide unified clarity and bridge gaps between parallel threads of research.

2. Saliency Maps

Given the apparent inaccessibility of deep neural network-based models, visualizing their inner workings has taken center stage in the machine learning community. But unlike in other more straightforward recognition tasks such as face recognition or car detection, understanding the generated predictions for more abstract tasks such as natural image categorization is much more difficult. This is, in part, due to the widely varying visual appearances of images belonging to a specific category. Above and beyond being just computational black-boxes that come up with seemingly mystical predictions for esoteric datasets, computer programs that deploy neural networks process and evaluate thousands of parameters in service of their task. Instead of looking into each of these parameters, saliency mapping attempts to understand the forward pass mechanism by which a feature is produced, and additionally turns the process inside out, associating the generated feature map with producing parts of the input image. This postulate that the relationship between the input image and the feature map is an inside-out version of creating the feature map both ways leads to a very powerful tool for visualization, reveals the workings of a neural network, and the generated feature maps have been surprisingly demonstrative, and surprisingly beautiful.

We show specific examples of using saliency methods such as Backpropagation, Guided Backpropagation, Deconvolution, and Deep Learning Textures in our work on Brain Tumors MRI Images, Melanoma Dermoscopic Images, Diabetic Retinopathy Fundus Images on RGB Images, Natural Images, and Text-To-Image Synthesis. Saliency Maps reveal the use of color in the Feature Creation process, show the localization of a significant brain tumor, neural network detection failures, instances of incorrect input data segmentation, and the salience of features omitted by Deep Learning Texture synthesis.

3. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) provides visual explanations for models trained with different objectives such as classification, detection, or segmentation. The core idea is to generate a heatmap in the input image to highlight the regions relevant to the model prediction. Equivalent to the traditional class activation mapping, Grad-CAM uses the gradients flowing into the last convolutional layer to produce the heatmaps. The gradient information reveals which regions in the image have affected the predicted class score the most, allowing for

multi-class categorization by simply weighting the contribution of each class in the linear combination [6,7].

Grad-CAM was applied to explain a neural network that predicts a patient's probability of adverse post-operative events. Different internal components within the neural network were analyzed by visualizing the framework along with the Grad-CAM output heatmaps. The Grad-CAM was visualized along the network layers to highlight the features learned from the model at different layers; this approach reveals how the model is learning. The model learned various useful features at different layers; certain models the fluorescence intensity signals attributed to the A, C, G, T bases, whereas the model being visualized appears to detect patterns at the time level in the data and there is an increased magnitude of importance of certain patterns along with the model prediction confidence values. This analysis also explains the location of the sparse filters [2,8-10]. The filtered signals of the Grad-CAM regions along the layer are then backtracked to highlight the regions in the original traces; these filters use only a subset of the bases to make the prediction, as evidenced by the absence of certain bases in the filtered section.

4. SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) unifies several previous approaches into a single framework based on the classic Shapley values concept from cooperative game theory. Shapley values are a method for fairly distributing a value among participants that have collaborated to generate that value. We will first explore how this approach to measuring feature importance arises naturally in the context of a whitebox model. Let's assume we were using a predictive model that consisted of a single number—for example, meant to estimate the average price of children's shoes in the US. However, since we had additional information available for each particular shoe in the dataset, we built a much more complex predictive model to improve the quality of each prediction [1,11-12]. Each new prediction is likely to be closer to the actual value than the simple model, and the difference between these two predictions represents the "value" generated by the predictive model for that input. Since we have a very low-dimensional model for the shoe price, we can "switch on" a different set of parameters for each shoe price prediction. We would like to distribute the credit for the predictive desertion among all the parameters, since they all had to collaborate. The contribution of each parameter at a particular prediction would be the difference between both predictions multiplied by the proportion of the training dataset on which this parameter was relevant.

In this explanation, we assumed a very simplified case, with a single parameter and linear varying predictions. Expanding this concept to a case with many parameters and

considering every possible distribution of the "credit" would yield us the classic Shapley value estimation [13-15]. SHAP directly generalizes this process, offering an intuitive understanding of how Shapley values apply to model-agnostic feature attribution. As a special case, SHAP allows us to divide the credit for a specific claim along the parameters in a more general class of cooperative models that includes Bayesian average models, exponential family models, finite mixture models and feature models. With this methodology in hand, we introduce the different ways of approximating the computation of Shapley values for specific models, allowing SHAP to function within a wide variety of techniques and output clear and intuitive predictions.

5. Counterfactual Explanations

Counterfactual explanations attempt to answer questions of the type "What could I have done to obtain a different outcome?" They are widely used to explain the decisions made by nondiscriminative classifiers, which predict an outcome using a true score or risk model that is calibrated against a reference population [16,17]. What the model does is to say how likely it is to observe the predicted outcome for someone with the given observable characteristics but does not hint at the probability of the actual individual. By definition, a counterfactual does not exist for someone whose outcome was predicted correctly: "For that individual, the model is predictive and accurate. It would be odd for them to go to their bank if they have a loan that is unlikely to be repaid." Researchers are divided on whether transferability allows for the explanation of single cases, why the first statement implies that transferability allows for broader statements than the second one, and why incorporating the observer perspective, when the model is assumed to be accurate, may help resolve the discrepancy.

Counterfactual explanations for people predicted to experience a negative outcome can be useful to design loss-mitigation strategies. By identifying individual characteristics that would have changed the decision for the risk forecasting model expressed, the model may help specify what can be altered to have a different predicted outcome and eliminate the proliferation of a specific undesirable cohort. More precisely, in the rectifying perspective scenario, counterfactual explanations could help identify which factors can be modified, as to rank and prioritize, or suggest risk-influencing attributes to increase the chances of a better outcome on a future occasion.

6. Case-Based Reasoning in AI

Case-based reasoning (CBR) is a form of AI that identifies solutions to new problems by reusing solutions from previous similar problems. This approach was originally inspired by principles of human cognitive processes in dealing with new problems and

by the research of the laws of analogy. The first CBR models were indeed cognitive models, and many systems were developed in such diverse domains as user assistance programs for software debugging, medical diagnosis, training, educational technology, case management, law and legal reasoning, prediction, etc [12,18-20]. At the same time, however, CBR systems incorporated a richness and sophistication of design and implementation, employing methods, techniques, and technologies for many different AI areas that gave CBR a status of a multidisciplinary field. In many domains, the types of problems to be solved or the conditions of the problem-solving environment do not lend themselves well to representation or preparation of knowledge through general or specific rules. The solution appropriate for only a particular case is not easily generalizable to cases in the same domain or problem area, as is the case when rules are used. Theoretically grounded principles for knowledge representation and knowledge acquisition for rule-based systems, for knowledge representation and problem-solving capabilities of different levels of competence, or for automatic generalization and specialization of knowledge are the exception rather than the rule [21-23]. Currently, the different pathways that exist between the AI subfields of CBR, neural networks, fuzzy logic, agent-based systems, multi-agent systems, genetic algorithms, systems of systems, and semantic web technologies are even more lively and fertile than they were previously. Each of them contributes to the design and implementation of new truly intelligent information systems that integrate the strengths of numerous AI methods and techniques [24,25]. Therefore, the combination of CBR with other AI paradigms and techniques is generating new, innovative, and sophisticated approaches and engine components that contribute towards the development of more useful, both in practice and in theory, AI knowledge-based systems.

7. Rule Extraction from Deep Models

Although rule-based models are typically among the simplest and least flexible for learning, they can be very powerful. A set of rules may effectively capture relationships within the input space, while remaining computationally efficient. Moreover, such rules are easily interpretable for end-users. It is thus not surprising that many in the high-stakes decision-making spaces desire for the machine learning systems they utilize to have similar levels of interpretability. Unfortunately, the temptation of predictive accuracy afforded by deep learning is often too strong to resist when performance is needed but cannot be afforded. That is why it is so crucial to investigate solutions that would allow users to extract at least some of the interpretability benefits of lower-capacity models from deep systems that lack any of the nicely structured representations that traditional machine learning systems produce.

Some of the earlier proposals for rule extraction were developed to produce low-complexity surrogate models from neural networks to predict their outputs on a computed sample dataset [26-28]. Although this kind of post-hoc explanation system is definitely useful in that it can provide insights regarding mode behavior or certain part of the input space, it is also limited by the rule sets being learned: they can only capture specificities of the modeled function in relation to the learned distribution. The advancements in deep neural networks using unsupervised and inductive methods primarily focused on using the individual building blocks of complex architectures like convolutional neural networks as explanatory parts of speech. These pieces of knowledge are indeed very expressive, and can often allow experts to provide detailed semantic descriptions about what certain parts of the model are concerned with.

8. Physician-AI Collaboration

Collaboration between physicians and AI has emerged as a productive and beneficial factor affecting the diverse domains of medicine. Proponents of AI-assisted solutions argue that they can outperform unassisted physicians by exhibiting super-human capabilities. Such a proposition is impossible to gauge in clinical use since, by design, patients are rarely treated without supervision [29-31]. Rightly assuming that there are benefits to both parties, a more sensible focus should be on the productive collaboration that AI tools enable. Fundamentally, machines and humans operate for distinct evolutionary reasons. Each has its strengths and weaknesses. In particular, the parallel pattern recognition of AI is well-suited for digitized data in large amounts while uncertainty investigation is by far the bigger strength of any physician. AI should therefore assist specific compartments of the diagnostic journey in terms of speed and efficiency while humans challenge the general diagnosis and interpretation or paradigm novel findings, as the patient's experience is not part of the machine approach. Translating these two ideas into actionable guidelines in everyday practice has turned out to be a serious problem, especially in radiology. Good partner behavior has been neglected in both clinical and research implementation. In a typical paradigm for mediating, radiology AI will implement pre-selection of possible findings enabling the radiologist or clinician to focus on a smaller area of interest. Once the AI proposes the result to be reported, it is again in the hands of the radiologist or clinician to discuss and either adopt or challenge the input. These decisions require familiarity with the use of specific tools as more have emerged that address specific tasks independently without collaborating. For example, a lesion exclusion tool can have detrimental effects on the decision process once a lesion is triggered but may greatly speed up the examination once lesions of no interest are identified.

9. Building Trust in AI Systems

Research on trust in human interaction has a long and rich history, beginning with the research investigating factors that support cooperation and altruistic behavior through to more recent work on the trust that human users place in automated systems to act faithfully and reliably in cooperation with human users. It was noted that the social biases that humans exhibit in interacting with other humans also extend, in some circumstances, to automated systems. It is observed that humans can be sufficiently influenced and persuaded to trust AI systems that they behave inappropriately, leading to misuse of an AI system. There is also work on how a relative lack of trust may inhibit the adoption of robotic and AI systems, particularly in safety-critical settings such as aviation and healthcare [3,32,33]. Early research on human-automation interaction supports the view that AI systems should be trusted enough, though not too much to ensure they are accepted by human users. Chatbots and self-learning AI systems can exhibit strange and undisciplined behavior, reducing trust by virtue of their errors or inability to explain their reasoning processes.

Research on interpretability in AI systems is particularly salient when examining the use of AI in clinical settings. The potential benefits of AI in clinical systems must be balanced with the ethical and safety concerns regarding the use of AI to supplement decision making in healthcare and to replace human decision makers in limited but important situations. Models that cannot be trusted or do not yet evidence safe levels of trust are unsuitable for high-stakes clinical environments. Furthermore, clinical users are likely only to accept the results of AI that augment or replace human decision-making if the AI is interpretable – for reasons that go beyond mere curiosity.

10. Challenges in Interpretability

Despite the growing importance of interpretability in AI systems, several challenges remain regarding the practical design and implementation of interpretable CML systems. One prominent issue is developing accurate or useful interpretations of the models [4,34-36]. Defining what is meant by interpretation and the criteria by which the quality of an interpretation is evaluated are often subtly different for users working in different domains or specialties. This introduces variance and uncertainty regarding the degree to which different AI models are able to justifiably assist clinical users in their decision-making processes. Addressing this challenge often means AI developers must work closely with stakeholders in a particular domain to evaluate what formulations of interpretability work best for that application.

Another interesting challenge to explore is the often debated trade-off between model performance and interpretability. Relatively simple models such as logistic regression models are inherently interpretable but may fail to capture complex relationships in the

data, leading to poor performance [37-41]. Conversely, complex models such as deep neural networks have the potential to accurately model even highly complex relationships in data but are difficult or even impossible to interpret. As a reconciliation attempt, methods for model-agnostic explanations can create human-readable interpretations for overly complex models but often require simplifications that introduce additional uncertainty into the model outputs. Thus, it is crucial to balance model accuracy with the need for interpretability when developing evaluation procedures.

11. Ethical Considerations

When systems are developed for deployment in the clinical domain, their contribution to social good must be realized. The ethical considerations of research in clinical AI address issues in transitioning from the discovery of useful machine learning solutions to a deployed tool that is ultimately the best option for solving a clinical problem [3,5]. Choices made during development have consequences for how likely a system is to realize its promise after implementation. Ethical concerns center on these choices, including whether to skip stages where interpretability can be examined and emphasized, a lack of prior clinical and real-world data for guiding design choices in model training, management of expectations around performance, inadequate workup of how to guide clinicians in understanding informally how systems function, and inadequate commentary on diversity and representativeness of training data.

Researchers and developers are frequently placed under pressure to train models that perform with high numbers for generic performance metrics. Being known for groundbreaking performance on a wide range of tasks has its own allure, yet such increased exposure can have unforeseen consequences. When research teams focus inward on these metrics as the bar for care transformation, they overlook the reality that, while predictive performance is an important metric, it is not sufficient at deployment. Lack of discussion of issues of model interpretability, data quality, and sample diversity can lead to hidden biases in our medical AI designs. These biases can amplify known biases in health care deferral practices.

12. Comparative Analysis of Interpretability Techniques

To help researchers select the best appropriate interpretability techniques, various taxonomies have been proposed to analyze the available options. One of the best known taxonomies of interpretability techniques considers three dimensions to describe existing interpretability techniques: the level of interpretability, explanation type, model characteristics, and thus they classify interpretability techniques along those dimensions. The first dimension, level of interpretability, refers to the fact that model interpretability can arise at different levels. They consider a level, which refers

to properties such as the concept they address, or the input dataset size they scale to, or the way they combine to produce an explanation. The second level refers to the way the interpretability mechanism produces an explanation. The Explanation Model is the concept that they define at that level. The output of the interpretability technique is an approximation of the model guilty of observing its input and output on a set.

Interpretability techniques are typically more efficient at producing explanations for a reduced number of inputs and outputs. Hence, the local explanation that they produce applies to input samples, while global explanation for a model focuses on the input space. The last factor that they consider is model characteristics, which can range from factors related to the data to property, and basically that being exploited by the communication model and comparison criteria is key to understand what explanation any model is best suited for. A global explanation can yield an ambiguous explanation when the model is evaluated on a number of samples. If the samples registered out an entire size of audiovisual input, watching would be ambiguous. Local explanation for this model would normally share that to a reduced number of samples the local axes might be too long, while the global instruction model might compare against sets of arbitrary size. However, if these properties are too theoretical and not ideally related to the model, both types of explanation could work well.

13. User-Centric Design in Clinical AI

Clinical AI systems are often utilized for high-stakes, individual decision-making based on predictions of complex patient-specific scenarios that heavily influence health outcomes. Moreover, clinicians interact with models through complex, multi-step decision pathways that are also particularly nuanced [6,9]. This stands in stark contrast to many standard ML applications, where usage often first occurs with aggregate decisions to identify distinct clusters in data, or a single-step decision. Consequently, the limited goals of interpretability in ML may not be sufficient and effective clinical AI systems would benefit from translation of usability principles from HCI. User-centered design techniques for clinical settings can help researchers better comprehend how stakeholders use predictive models and the decisions they make based on the models' predictions. This not only enables more effective, interpretable representations of model predictions, but also can directly inform more effective development of the predictive model itself. These concepts could also be used to guide the design of public and regulatory policy around clinical AIs.

By incorporating user feedback, designers can address important elements of the user experience. For example, alternative output formats and visual designs may improve users' access to hidden states or user affordances, such as showing model uncertainty or creating an easy comparison with related populations. When such aspects are

designed thoughtfully based on a good understanding of their effect on interpretability and usability, confidence in model defense and repeated use may improve.

14. Impact of Interpretability on Clinical Outcomes

Several studies have assessed the impact of model interpretability on actual clinical judgment or patient healthcare outcomes. In a simulated setting, a group of researchers asked dermatologists to evaluate clinical vignettes of skin cancer lesions, either on their own or paired with model predictions from two systems—one with a simple model that only yielded a class label, and another with an interpretable model that highlighted regions within the images that contributed to the prediction. The findings indicated that expert dermatologists were less accurate when asked to evaluate the lesions with their paired model predictions than without—and this drop in accuracy was significantly steeper for those paired with the interpretable model than with the less interpretable one. In a follow-up study, a similar setup was used to compare two interpretable models. The conclusion was that "explanations in AI-assisted dermatology models can redirect dermatologist attention and may hinder diagnostic accuracy." In another study, the impact of different types of explanation approaches in two real clinical settings was examined: a text-based prediction and its explanation using two different explanation approaches were embedded into a clinical hospital system, allowing healthcare professionals to perform a simulated triage application using these embedded explanations. It was confirmed that both types of prediction explanation (albeit using different methods) have an equally negative impact when compared to the class label alone.

More recently, a real-time, interactable local interpretation method using Shapley values was evaluated on a clinical cohort of patients and the interpretation results were directly transferred into a real clinical deployment of a prediction task that helped doctors identify at-risk patients for major adverse cardiovascular events. It was found that enabling Shapley interaction reduced doctors' intervention costs while improving MACE risk identification in type 2 diabetes patients.

15. Future Directions in AI Interpretability

This chapter surveys the present research directions in AI interpretability, organized in the report card format by three central questions: (A) What does interpretability mean in AI? (B) What is the interpretability gap in neural networks? (C) What are common techniques to interpret AI systems? In this chatbot era of AI interpretability, we find ourselves at an existential moment. The winds of fortune are at our back, but what should we do? Can we afford to train more interpretable AI models, i.e., models with sparse parameters inspired from neuroscience? Or perhaps continue training larger transformer models passed first on flocking, then on Linnaean taxonomy, and are now

setting course for spoken language? The current AI interpretability research landscape is a kids draw: there are more questions than answers. What does interpretability mean in AI? What is the interpretability gap in neural networks? What are the common techniques to interpret AI systems? Do we need AI interpretability during training or only at inference time? Does AI interpretability challenge the generality of the model? Or does it push for performance at the level of the current checkpoint? Finally, and most importantly, when is an AI system interpretable enough? We admit it is bad practice to put open questions on a road sign. This work is by no means a survey of the current research in interpretability, although we cite a few representative contributions. Rather, we want to emphasize the relationships between the open questions in interpretability and the contribution of specific papers. We hope that doing so will let the reader find their own way through the dense forest of interpretability work, and make connections with the interpretability problems they are tackling.

16. Case Studies in Clinical Settings

Investigating the interpretability of AI systems in clinical settings is important, not just because it will impact patient outcomes in a unique way, but also because there is already a large experimental record of the difficulties of physicians in understanding the outputs of AI systems in the clinic. Traditionally, the clinical record has contained an abundance of data about clinician decision-making but little about the actual character of the human-model interaction. With the advent of clinical AI systems, which monetize worse patient outcomes, the ethical stakes have increased.

Here we present work to untangle the human-model interaction - specifically, the interpretability of AI systems in real clinical use around the world. We also discuss how these attempted answers could help future developments in clinical settings. Our work ranges from studying how errors generated by AI systems propagate to decisions made by clinicians to how the predictive performance of deep neural networks affect uncertainty quantification and human trust in a healthcare setting. It is still at a preliminary stage and will be further developed in the future, but we nevertheless hope to impact the features of future clinical AI systems as well as our understanding of user-model interaction in a clinical AI context.

By relying on a selection of case studies to put into context our discussions, we believe we can bring a better understanding of the interaction. Nevertheless, as we detail throughout this section, systematically addressing the explainability of model predictions in clinical practice will prove challenging as practical considerations and the real-world environment are key modulating factors on the need for understanding these model predictions.

17. Regulatory Perspectives on AI Interpretability

AI interpretability is increasingly seen by regulators as a critical component for assurance of safety and efficacy when AI is used in high risk domains like healthcare. This chapter reviews the guidelines and opinions that regulators have issued regarding AI interpretability, and argues that while the specifics differ, there is alignment of high level expectations on industry between different regulatory bodies and across different jurisdictions. The chapter then looks towards the future, questioning whether the traditional regulatory approach is enough to address the problems presented by clinical AI, especially in knowledge domains that are poorly specified by current regulatory frameworks.

In recent years, numerous government and inter-governmental agencies and committees have published opinions and guidelines on the subject of AI and ML Accuracy and Applicability and on what developers and deployers of these systems should do to steer machine learning in the right direction. Some of these efforts are perhaps best categorized as "general" guidance. Other efforts are official recommendations of specific interest for healthcare and/or life sciences. And others are efforts by groups of researchers in the machine learning or clinical domains headed by publishers or academic institutions.

18. Integration of Interpretability in Clinical Workflows

The interpretability of AI systems in healthcare is crucial when it comes to ensuring that the promises made during the design and development phases are met during deployment and use. Furthermore, it is essential to convince various stakeholders—such as clinicians, patients, regulators, or funding bodies—that there is added value in adopting these new technologies, as opposed to continuing to work with traditional methods. Over the last two decades, a significant body of work has been conducted on several different explanations targeting diverse stakeholders with diverse needs. Those approaches include local versus global explanations for clinicians or explaining predictions about certain patients on a one-by-one basis; explanations targeting reasoning and evidence or certainty; visual versus textual explanations; and many more. Some of these principles are mutually exclusive. Explanations aiming to recall memory and draw global conclusions will be difficult to combine with those explaining only one patient at a time. Furthermore, scarcely any of these systems are integrated in the workflows used on a day-to-day basis by clinicians, in particular when it comes to clinical reasoning tasks. Usually, such explanations thus have to rely on either clinical reports or special research software not intended to be used in direct contact with patients, in real clinical scenarios.

The situation is different for diagnostic sections of clinical reports stemming from automated processes. While there is a great deal of interest toward integrating generative large language models in the workflow, in an attempt to make them more reliable, their output is still less reliable than that produced by fine-tuned disease classifiers. That of course poses a double challenge: making the output of the ML-generated diagnostic report reliable and ensuring that physicians and patients pay attention to it so that any doubt is cleared before the patient leaves the ward. Another way to entangle explainability with clinical workflow is to include interpretability techniques for ML-augmented diagnostics in the study of abnormal cases, similar to the classification for cancer staging. Such investigations are usually done on specialized units stacking difficult cases, in order to share knowledge among several specialists.

19. Training Healthcare Professionals on AI Tools

It is crucial to fill the gap between the technological aspect of clinical AI and the actual use of it in clinical environments to increase the users' trust in AI systems. There are two factors that can support a collaborative approach to AI, improving users' trust: clinicians' understanding of how AI works and a simple interface for asking for explanations for AI decisions. The experience with clinical AI-embedded medical devices often is described as more educational than that with other clinical devices due to the lack of formal training. Proper training of users is the best option to diminish their uncertainties on clinical AI. However, currently, healthcare professionals are poorly trained about AI, as the pedagogical framework is lacking and most training sessions are carried out only in a one-off manner and often these are just informal discussions at departments or labs.

In order to train healthcare professionals in the workings of AI clinical systems, it is better to adopt a tailored fashion to the issue than to provide a monolithic digital health literacy. This is because the need for training can vary not only in the type of healthcare provider but also according to their clinical specialty or sub-specialty. For example, it is possible that one healthcare profession or specialty might be excluded from the initial target, resulting in implementation models and training procedures that are not suitable for that specific reality. Therefore, the implementation of clinical AI systems cannot rely only on a basic level of digital health literacy, such as the ability to understand and use easy AI tools, but should also consider a specialized approach. In turn, this point leads to consider how the evaluation of more advanced abilities can be structured and a possible differentiation between the general and specialized levels of digital health literacy.

20. Feedback Mechanisms for Improvement

Feedback mechanisms are vital processes for encoding system-level experiences into future designs of Clinical AI Systems. Feedback on AI system accuracy is important, and in used systems, ground truth data is produced for retrospective accuracy checks. Capture of why-why data, which records justifications given by the Clinical AI System, clarified by human auditors correcting incorrect outputs, may reveal human- and machine-based sources of error. These reports fed back into system designs may allow for retraining of Clinical AI Systems for lower output error, or human worker groups may be trained by revealed group-level trends to rectify why-why reasoning errors viewed as common across diverse cases and contexts. Altogether, capture of general patterns for problems with the Clinical AI System output may allow for diverse improvements in accuracy. Of particular importance are interplay loops, where the human and machine components of a hybrid system allow for retroactive capture of the human and/or hybrid system output. Positive demand from humans for A-assisted work, plus capture of specific cases in which A-output was incorrect but human output was correct, and demand from humans by capturing cases where Human output was incorrect, but A-assisted output was correct, allow the AI and Human components to improve from informal feedback. Human augmentation systems that receive informal and ongoing feedback from workers have been observed in industry to be far more effective than systems without such feedback loops. Solutions echoing the humility and complementarity ideals of human augmentation applied to iterative feedback loops hold strong promise.

21. Patient Perspectives on AI Interpretability

As co-developers of healthcare technology, it is essential that patients are involved in the design process for interpretable AI systems that use the patient information for driving decision making. AI interpretability does not take on a single meaning for all users. For example, physicians may define interpretability as the ability to explain the decision-making process, whilst customers may seek additional information justifications, as well as persistent systems that enable exploratory data analysis. Moreover, both patient expectation and provider need depend on the context of use, and the service quality of the biomedical AI service. AI interpretability has been addressed mostly from the provider's perspective. In this short section, we discuss opportunities and needs arising from the patients' perspective on AI interpretability as a proposition for further research collaborations between AI developers and the clinical experts. For AI to have the desired impact on healthcare delivery, it is important to understand that all stakeholders expect different things from AI interpretability. They would like to obtain answers to different questions when interpreting a model trained using their existing patient records or gene sequence data.

Such diverse interpretability wishes stem from the fact that existing interpretability methods primarily explain the effect of individual or a small number of features on the model output, which AI users attribute to inquiry different questions about the model. A significant challenge is determining what is actually being captured by these individual or small feature results. Providing such interpretation cannot be done in isolation; the healthcare stakeholders, whose wishes guided the design of the model should also take a part in the recovery and interpretation from any AI model explanation tool, rather than trusting the interpretations made by others.

22. Technological Advances in AI Interpretability

Most methods for AI interpretability draw on established relations between features and a model input or between a model and its predictions. Since these relations only hold in a limited and rather artificial sense, any explanations produced must be taken with a grain of salt, because they often fail to faithfully represent how a human would arrive at the same decision. Due to these difficulties, many of the early works on interpretable AI have focused on decision trees, logistic regression, linear SVMs, and other methods in this family. While these models are globally interpretable by virtue of their simplicity, they often produce unreliable predictions on real-world datasets.

This is where the developments we mentioned above come in. As feature-representation learning matured, these methods began to augment the more traditional models, attempting to combine their sampling properties with the representational power of deep networks. From this perspective, a salience map for an input face image should produce a spatial distribution over the image such that regions that a model relies on for predicting a human’s identity have the highest values in the salience map. Rather than learning a distribution with that property directly, the more conventional approaches approximate it with a supervised model that operates on salience maps as labels. This kind of model can only be trained on images for which we already know the strong features that make the difference.

23. The Role of Data Quality in Interpretability

It has long been claimed that “garbage in, garbage out.” How data is selected, pre-processed, and represented shapes any model's ability to adequately project inputs to an output space, thereby defining the limits of predictive performance and, consequently, interpretability. For clinical practice, good quality data is fundamental as it drives the decision-making process and clinical pathways. Unsurprisingly, imperfect data can lead to model predictions and interpretations that are at best misleading and at worst harmful. In addition to data quality issues within AI predictions, display systems may further exacerbate perceptions of bad quality. Predictive models offered within a clinician's workflow need to be simple, trustworthy, and verifiable. Interpretability in

medical AI is not just about being able to explain a model's mapping of inputs to outputs; it is also concerned with the quality and uncertainty of the underlying data and which is closely linked to the builders' knowledge of the data domain. Just as a biochemist or radiologist knows their domain intimately, developers of AI tools should also be aware of the implications of domain knowledge on data quality and the importance of its intersection with a model's predictability when judging accuracy.

The attention given to uncertainty within clinical guidelines, clinical decision algorithms, and diagnostic accuracy needs to be paralleled by AI visualisations, otherwise, the invalidity of those tools can be exacerbated. A model may give a sufficiently good mapping of a space of inputs to outputs, however, it says nothing about the confidence of that model; and yet assuredness when predicting – especially when decision thresholds are set too low – can be hugely damaging. The intersection of model uncertainty and quality is expanding and rich areas for exploration for applied AI in medicine.

24. Cross-Disciplinary Insights into AI Interpretability

Generally, interpretability and explainability are research disciplines that cover alerting stakeholders against poor AI system predictions and increasing users' trust in a system. Although the approaches developed in these diverse research areas look quite different, these disciplines differ in their definitions of "explanation" and purpose of explanation, and solutions proposed to engineer machine learning models with such capabilities look also quite different. After first summarizing the differing definitions and roles of explanation - why explanations are needed in decision making - given by different ways of viewing systems as black-boxes, scientific instruments, or security silos as well as the differing approaches taken by various application domains, we will introduce how diverse and, in some cases, contrary recommendations provided by these disciplines may be reconciled.

Because AI system users may need explanations for different purposes, this diversifies the requirements for explainable AI. For example, a user may want to understand how the AI system performs prediction or reasoning, and be confronted by an unexpected outcome, in which case they will want to understand why the current input led to the observed outcome - with the goals of debugging the system to improve its reliability, identifying situations where the current input distribution differs from the training input distribution to explain possible poor performance, or reasoning on the basis of the predicted output during usage. In the case of high-stakes decision making based on explanations provided, the user may require rigorous soundness or correctness guarantees to check both the quality of the extrapolated output and whether the model reflects a causal relationship between input variables and output.

25. Impact of Interpretability on AI Adoption

The success of clinical AI systems is intimately related to how and whether they are adopted, and subsequently integrated, within clinical practice. The adoption of clinical AI systems appears to be an extremely complicated problem that has been unrealistically simplified by the framing of clinical AI systems as uncontroversially beneficial tools. In reality, AI systems are not simply tools for increasing efficiency. They are a far more complex category of high-impact decision-support systems that can potentially affect patient care, training and incentivization of clinicians, power dynamics regarding the governance of healthcare, and the essential experience of being a patient. A conversation regarding the adoption of clinical AI systems is then not simply a matter of conveying more accurate information or producing better evidence for the cost-benefit analysis of the clinical AI system. Adoption concerns the values and goals of the stakeholders in relation to the goals, risk classification, and value alignment of the AI system, as well as to its role and importance in the overall responsible governance of healthcare.

The importance of interpretation in these discussions is widely acknowledged, with some calling for an a priori requirement of interpretability that would be a necessary condition for medical device approval. However, whether and how interpretability affects the adoption of clinical AI systems has not been investigated in a systematic fashion. The findings of studies in HCI and education might not be directly applicable in this domain since the interplay of values and concerns is different, and the consequences of designs and misalignments are on a different scale. We describe the results of a study that explores the impact of interpretability and interpretability support when people consider an AI system for use in a medical context, and also when they project that context into an atypical domain.

26. Summary of Key Findings

In Section 1.2 we introduced this work and laid out the case for why interpretability is crucial to the success of AI systems deployed in health care. Namely, AI will be most successful in health care when we can trust the decisions made by these systems. To earn trust and facilitate successful deployment, we must understand their decision processes. Trust derived from interpretability will also facilitate model improvement, development of novel algorithms, and insight into clinical data. However, design and development of interpretable clinical AI models is a complex, multidisciplinary challenge. In Section XXX we outlined key principles that researchers should be guided by when conducting this multidisciplinary work. Primarily, we must involve stakeholders, including patients, in the AI design process. We also emphasized that while future systems may be able to explain decisions made by opaque black-boxes,

we should focus on building intrinsically interpretable models. And finally, while we argue that explanations can be valuable, we must remember their limitations, especially the potential for explanation-based harm.

As described throughout the book, there has been much progress in both understanding clinical AI model behavior and making models more interpretable. To enable informed future research decisions, in this section we briefly summarize some of the major findings highlighted in previous sections. There is growing evidence that AI model performance is correlated with clinical importance and is helpful for earning trust. Using intrinsically interpretable models that produce trustworthy predictions and decision information is a potentially superior approach as compared to inserting post hoc explanation mechanisms into a more opaque model. However, the latter approach is still valuable and at the very least will help us better understand black-box models. These approaches should continue to be combined to electrify our extrapolative view into the predictive behavior of models. Furthermore, as we improve on our ability to accurately model the clinical data, we may very well elucidate existing statistical, and perhaps even clinical, relationships within the data.

27. Conclusion

Developing interpretable clinical AI systems is not only important but also necessary to ensure that stakes associated with incorrect AI model predictions do not result in irreversible consequences. Achieving AI model interpretability, however, cannot be viewed as a one-size-fits-all paradigm. The diversity in design, stakeholder, and use-case considerations necessitate diverse interpretability paradigms to develop interpretable AI systems that are not only useful in a clinical context but achieve the objectives of the stakeholders that interact with them.

We argue the AI models used in healthcare should explore an explicit and direct relation with model interpretability. The anticipated consequences associated with incorrect predictions should be the primary factor in the design of AI model frameworks, selection of model-building constituents, and the envisaged interaction of stakeholders with AI models. By understanding the reasons behind the adoption of an AI model, a specific decision can be made about what should be interpretable – the model, its mapping from input to output, or the relationship between either the model or its output and the stakeholders. Our viewpoint is that adopting a clear stakeholder-centric approach is likely to facilitate the desired impact on healthcare.

References

[1] Mathew DE, Ebem DU, Ikegwu AC, Ukeoma PE, Dibiaezue NF. Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Processing Letters. 2025 Feb 7;57(1):16.

[2] Gliner V, Levy I, Tsutsui K, Acha MR, Schliamser J, Schuster A, Yaniv Y. Clinically meaningful interpretability of an AI model for ECG classification. NPJ Digital Medicine. 2025 Feb 17;8(1):109.

[3] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.

[4] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.

[5] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.

[6] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.

[7] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.

[8] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.

[9] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.

[10] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.

[11] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.

[12] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.

[13] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.

[14] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).

[15] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing

healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[16] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[17] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[18] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[19] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[20] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.
[21] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.
[22] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.
[23] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.
[24] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.
[25] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[26] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[27] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[28] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[29] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[30] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[31] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[32] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.
[33] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.

[34] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.
[35] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.
[36] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.
[37] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.
[38] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.
[39] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[40] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.
[41] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 6: Artificial Intelligence-Driven Clinical Decision Support Systems (CDSS)

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Clinical Decision Support Systems

Since the dawn of mankind, there has always been some type of decision support in every activity we carry out. From the early days of families to large organizations with formalized hierarchies, opinions, advice, and suggestions come in order to improve the quality of decisions. Among all human activities is the decision-making process regarding health problems which has always existed. With the pre-scientific, mythical and magical control of illness and health, learned by experience shamans, priests and wise men, they made some diagnosis, explanation and therapeutic prescription on how to deal with problems relating to health for a specific person or group [1,2]. These activities were becoming more clinical and scientific and during the academic age and the Renaissance; university academic groups began to accumulate knowledge on the functioning of the body [3-5]. In the following centuries, these groups systematized that knowledge into rules for making decisions related to specific diseases. This was how decision support in the health area evolved naturally. Traditionally, decision support in the health area was carried out on a face-to-face, personal, and unique basis without the technology-supported standardization that we customarily apply in today's world. But as in decision-making in other activities, the passage of time and the evolution of technology and the information area has allowed the development of systems that automate and facilitate decision-making in all activities, including in health. In the health area, the first technological supports to facilitate decision-making were elaborated in the late 1960s and the 1980s. These systems, generally based on Artificial Intelligence or formal logic, are called Clinical Decision Support Systems. These systems are tools aimed at improving the quality of care through the availability of clinical knowledge that generates recommendations for decision-making in specific

clinical situations and carry out the processes of diagnosing, treating and monitoring patients.

2. Overview of AI in Healthcare

Artificial Intelligence (AI) is a computational approach designed to perform judgments normally associated with intelligence. Machine learning is a subset of AI that encompasses a variety of statistical methods utilized to derive predictive models capable of recognizing patterns occurring in data and constructing future observations [2,6]. The AI discipline covers a wide variety of methods, including expert systems, computer vision, natural language processing, and robotics. The impressive growth AI systems have experienced can, in great part, be attributed to their successful applications in various domains, culminating in healthcare.

The presence of AI in healthcare is increasingly becoming apparent in the form of patient risk stratifying, lab test ordering, clinical guideline development and updating, and diagnostic support and prognosis. The factors contributing to AI's increasing assistance in healthcare include: (1) widespread evidence that small Bayesian networks yield easily computable probabilities and demonstrate clinically relevant accuracy in the risk stratification of patients; (2) AI-based clinical decision support systems' capacity to independently create and update clinical guidelines; (3) the remarkable success AI systems have achieved for image interpretation tasks, such as the detection of diabetic retinopathy and other diseases using retinal fundus images; and (4) the recent impact of neural network models on the capacity of hospitals to collate clinical data and use natural language processing solutions to disclose essential clinical patterns hidden in unstructured data [7-9]. These examples illustrate that, while risks remain, the finalized proof of concept is in patient outcomes, and that assurance is therefore possible through the careful AI-based modification of patient management pathways and clinical pathways.

3. Prognostic Models

3.1. Definition and Importance

Prognostic models predict future outcomes without the influence of an intervention or treatment. Consequently, these models provide important insight into an outcome's natural history, aid in informing treatment decisions, gauge the response to interventions, and allow for the adjustment of prognostic risk when estimating the effect of treatment on a clinical outcome [10,11]. There are several aspects of an outcome's natural history that can be modeled; including the probability of an outcome occurring or not during an interval of time; the time until an outcome occurs or is censored; and, for certain outcomes, the magnitude of the outcome.

The vast majority of prognostic models are developed to either predict the probability of a clinical event or the time until an outcome occurs. With the explosion of various types of data in medicine, researchers have crafted many different methods for developing these models as well as a plethora of variants on existing methods that allow for added flexibility or consideration of tested assumptions to predict one of the three above-mentioned outcomes. Furthermore, as testing methods extend to larger and more complex data sources, the use of artificial intelligence and machine learning is becoming an increasingly more attractive option to model prognostic outcomes.

3.2. Types of Prognostic Models

Leaving aside the burgeoning field of machine learning for the moment, the models that have been utilized most often to predict binary outcomes include simple logistic regression models as well as, more recently, elastic net and Lasso penalized regression models, classification and regression trees, classification forests, Bayesian methods, and several flavors of support vector machines [12-14]. For continuous outcomes, researchers have primarily relied on linear regression models, Bayesian methods, recursively partitioned linear models regression, and support vector regression.

3.1. Definition and Importance

Prognostic models are formal representations of knowledge regarding time to event distributions. They summarize (or are based on) the information contained in the individual prognostic factors and their collective effect on time to event. Overall, prognostic models serve to estimate the time until an event occurs or derive the probability that the event occurs at a specific time point. However, despite being routinely used in clinical practice, prognostic models still do not get the attention warranted considering their potential impact on clinical decision making. Indeed, prognostic models are the basis for clinical decision support systems [3,15-17]. Using a knowledge-based systems framework, these systems are an emulation of clinical decision-making using a set of heuristics that lead to certain diagnosis and therapy outputs. In the case of prognostic systems rooted in parametric models, heuristics link the input prognostic factor information to hazard or event probability tables, enabling clinicians to predict event occurrence at known time points. Meta-analyses show prognostic systems to improve clinical outcomes and the quality of clinical practice, such as decisional conflict and accuracy.

The primary rationale for using prognostic systems is that their risk estimates are more accurate than the gut feeling risk estimates provided by the human clinical decision maker. Indeed, the performance of these systems is mainly determined by the uncertainty of the individual and cohort specific risk estimates. Since prediction diversity is higher and prediction usually more accurate in models than clinical

decision makers, sensitive and highly uncertain clinical predictions are best assessed by prognostic systems. These additional benefits may vindicate their usage despite the concern raised that the use of computer algorithms in clinical decision making may dehumanize the process.

3.2. Types of Prognostic Models

The ultimate aim of prognosis modeling is to learn a predictive function from the training data, i.e., to devise a model that relates the input characteristics, or features, more formally known as prognostic or predictive variables, to the outcome of interest [18-20]. Given the type of outcome variable, i.e., the variable to be predicted, models can be broadly categorized into three classes: survival or time-to-event models, outcome classification or risk prediction models, and response modeling or prediction-of-predictors models. Survival models predict the time until an event occurs given the prognostic variables, and also account for censoring, i.e., situations when the event of interest does not occur for some subjects during the study's period and the data, essentially, right-censor their follow-up. Example survival outcomes include time to death from breast cancer, time to the return of prostate cancer after treatment, or time to experience a first episode of thromboembolism in venous thromboembolism patients on anticoagulants. Risk-prediction models predict a binary outcome variable that reflects whether the event will occur or not [21-23]. These models do not consider time, nor account for censoring. Example outcome predictions include "will the patient die from breast cancer in the next five years?" and "will the patient return to the hospital within the next month?" Finally, prediction-of-predictors models aim to predict either the quantitative or the binary outcome variable showing the strongest association with the time-to-event outcome, which is itself a predictor of the outcome of interest. The response can be considered a summary of all subjects most likely to be predicted by the potential outcome model. Without any loss of generality, the first two classes will be used subsequently to present prognostic results from numerous application areas in biomedicine and for a range of medical conditions.

3.3. Case Studies and Applications

Although most research on clinical prognostic models deals with risk estimates for survival without events, many studies in various cancers report a wide array of different events after diagnosis and the need for danger estimates for the events of interest is growing. As a consequence of this increased need for clinical precision medicine, many studies also report on prediction modeling for pretherapeutic events such as a pathological complete response after, for example, neoadjuvant treatment in breast cancer. Models of this type help with treatment decisions and also with

counseling patients about the likelihood of achieving a pCR (an event leading to a better prognosis).

Finally, prediction models concerning treatment-related adverse events, such as therapy side effects, surgical complications, or treatment failures, i.e., not reaching the originally defined treatment goal, are also presented. The need for these models is not only one of statistical interest but allows for more personalized clinical decision making, shared decision making, and more accommodating patient information in clinical practice. Probably the oldest type of model for predicting the risk of treatment failure is the actuarial estimate developed for the surgery of cancer. In fact, those were the first so-called prognostic models for patients undergoing cancer treatment mentioned in the literature. Since then, many groups and consortia reported many different competing risks.

4. ICU Prediction Systems

Critical care medicine is perhaps the most constrained environment for clinicians attempting to provide patient care. Intensive care unit (ICU) patients are characterized by the need for constant monitoring and assistance, and the presence of severe physiological derangement and life-threatening illness [9,24,25]. One can thus imagine the natural appeal of systems-based approaches to the needs of the ICU - the presence of quantitatively abnormal physiology, continuous data streams, and a predilection toward pathological derangements that follow predictably wrong trajectories for particular disease processes all present conditions that are favorable for computerized intervention. However, the challenges posed similarly present large hurdles to the successful implementation of predictive algorithms.

ICU decision support systems in practice today excel at assisting clinicians in the interpretation of often disparate data streams from many patients and providing these data in easy visual formats, emphasizing intervals of derangement or potential future derangement. By contrast, true predictive algorithms use patient-specific data to estimate physiological states in the near future, reporting the most likely outcome for future time intervals, based on previously well-studied statistical correlations between variables and known clinical outcomes [26-28]. Various predictors have been developed covering a wide range of ICU outcomes, including renal failure, metabolic derangement, the development of sepsis and septic shock, extubation failure, re-intubation, shock recurrence, readmission to the ICU from the floor, and death and morbidity both during the index hospitalization and after hospital discharge.

4.1. Role of AI in ICU Settings

Predicting important clinical endpoint or clinical course of patients is essential in ICU settings. Previous studies have shown that although doctors are capable of using cues from patients conditions to predict important clinical endpoints, they are not at all times accurate in their predictions. Artificial Intelligence, especially machine learning algorithms, may be able to augment clinical decision making and help predict outcomes of patients more efficiently. IA-driven analysis of clinical data is of growing interest since it can potentially discover previously undetectable patterns in high dimensional clinical datasets [6,29-31]. Training predictive algorithms to create mathematical risk models can provide a useful addition to clinical decision making such as informing doctor-patient conversations or guiding management to address higher risks of complications. Furthermore, such risk models are increasingly being used in the ICU setting to alert clinical teams of significant complications.

Many algorithms and tools have been developed to predict a variety of different outcomes for ICU patients and often focus on short term physiological outcomes. Work has also been conducted to define deterioration beyond the confines of a specific clinical problem to include general mortality risk, such as mortality prediction tools. Within the realm of critical care, it is important to examine the clinical problems that are being predicted and whether a prediction tool based upon the model has a beneficial application in practice [32,33]. Predictive modeling has been previously described as one area of research that could fundamentally alter daily clinical practice in the ICU. In regard to state-of-the-art models, it has also been suggested that any prototype assess to improve quality, economic, or safety metrics before it is considered for integration with clinical decision support systems.

4.2. Predictive Algorithms and Their Effectiveness

A myriad of machine learning algorithms have been proposed to predict outcomes in an ICU population. Recently, transformer-based models have demonstrated good prediction performance, even outperforming the widely used LSTM models. Numerous commonly observed clinical events and parameters, including age and sex, diagnostic codes, operation codes, charted lab results, lab abnormalities, and glucose and blood pressure levels have indeed shown sufficient association with outcomes of interest when used in predictive algorithms [34-36]. In one recent study, these traditional clinical events and parameters were found to outperform more innovative time-series signals in 30-day mortality prediction in patients undergoing surgical procedures. Other models have shown similar performances when charted clinical events and lab data are utilized to predict varying ICU outcome parameters, including adverse post-operative events and hospital length of stay. To manage the exploding

volume of health data attributed to the growing digitalization of clinical environments, one of our works has focused on developing a high-performing development pipeline for future predictive algorithms.

Here, we successfully show that deep learning predictive algorithms can match and even outperform prevalent non-deep learning systems, without the need of complicated hyperparameter tuning and feature engineering and if not better, as good, when compared to other current data mining and machine learning systems. We perform the first scalability study to understand the effects of both the "curse of dimensionality" and label sparsity on predictive performance. Our method demonstrates why, when, and on what input data such predictive algorithms can scalable be developed to help clinicians triage and prioritize patients during their hospital stay at the earliest time and for any time interval. Moreover, these models could be generalized to facilitate soon-to-be-published deep learning-based CDSS to improve clinical workflows during such competitions and enable in-depth patient recovery while ceasing potential patient harms pre-, intra-, and postoperatively.

4.3. Challenges in Implementation

Although the sensitive, time-critical, and multi-modal tasks performed in open and closed loop scenarios at the intensive care unit (ICU) could benefit greatly from partially or fully autonomous AI systems, the combination of these complexities proves challenging. AI vigilance will be important in addition to the general ethical issues associated with high-level robot autonomy. In particular, ethical aspects specific to patient care should be considered, characterized by an unequal relationship between caregivers and patients [16,37-40]. These challenges are identified in existing literature on robot autonomy and ethical issues.

The prediction of patient deterioration in ICU settings allows providing timely therapy and avoiding negative outcomes such as unplanned intensive care unit admission or transfer to another level of care. To realize the advantages of early warning systems for serious deterioration, we will need to overcome many barriers to implementation. Barriers to implementation include unwillingness of clinical staff in the predicted event's investigation, fiscal disincentives for reducing surges in transfers from wards to ICU, and inadequate KPI definition. Some of these barriers will reduce the effectiveness of technology implementations, leading to a loss in confidence from which they will subsequently struggle to recover. In the same vein, inadequate stakeholder engagement, for instance in oversizing the use case scenarios or selecting suboptimal prediction horizons as deployment settings, will have detrimental effects on both the transferability and the later operational performance.

Despite the challenges, I see a clear advantage in intention-based predictive technology compared with currently used triage systems in wards and ICU. These systems to which the technology will be contrasted are inherently wanting in terms of specificity and are thus suboptimal already as clinical decision-support systems because of the risk. Decision-making that is uncertain should best be informed through probabilities and patient-specific reasons for their prediction, even if the quality of the latter falls short of unreasonable quality expectations.

5. Triage Systems

Triage is a natural domain for the application of CDSSs using AI. Emergency Department waiting rooms can be chaotic, such that experienced clinicians have difficulty developing a situational awareness regarding changes in patient condition to enable appropriate workflow management. Many patients within waiting rooms are acutely ill or developing critical conditions such that short-term assessments are not sufficient. Triage involves estimating need for intervention, acuity, and urgency, often seen as separate constructs [41-43]. Need for intervention and acuity are two of the constructs in the standard scoring systems used internationally. Triage uses basic data collection, primarily patient's presenting complaint, vital signs, and the clinician's "gut" feeling, which may involve rapid observational experience regarding similar presentations to manage the flow through emergency departments. The nature of the process, with both high-risk patients being assigned to the most urgent category and poor matching criteria for each category means that lack of precision can be acceptable. Modest improvements for patients who are particularly low acuity can have a large impact on overall throughput.

Hierarchical multinomial modeling has been used to augment the triage process for patients with characteristic conditions. Multiple AI enhanced triage-related systems have been investigated [44,45]. One system aims to quicken the process. The systems include an image-based system, a system based on the process of patient waiting and those utilizing the other inputs to the full triage system. Another system is aimed specifically at remote locations where mobile health is being utilized. Other systems combine mobile health with telehealth. One system utilizes telehealth interfaces but is aimed at improving ED capacity. This work illustrates the wide range of different designs possible within remote and on-site telehealth interfaces and protocols. One approach has utilized two different methods to enhance fewer resources in remote locations. The emergency clinicians may both remote assess a patient and impact the triage decision process.

5.1. AI-Enhanced Triage Processes

People seeking immediate medical assistance usually visit an emergency department (ED), where preliminary treatment is delayed while emergency-care physicians attend to same-day admissions with more severe conditions. Because not all emergency candidates suffer from critical illnesses, the handoff needs to efficiently prioritize the patients' assignment to clinical treatment, and this operation is usually performed with the support of a standardized triage tool. Through general signs, symptoms, and medical history during the first assessment, structured inputs are employed to categorize patients into corresponding severities to determine the urgency of their conditions, flag the non-urgent cases, and reduce flow to primary care clinics. Also acting as a bridge between patients' arrival and clinical evaluation, the triage process is important to patients' prognostication, emergency quality assessment, and clinical predictive modeling. Triage relies heavily on clinical expertise, though subjective judgment is inevitable; moreover, the demand for ED triage by physicians is not a realistic prospect because ED usage continues to increase steadily.

Aiming to calibrate emergency care by producing unbiased and objective triage predictions, artificial intelligence (AI) has the potential to assist clinicians in making the final decisions based on all the patients' characteristics. Characteristics of advanced computing infrastructures integrated with healthcare information systems enable AI to process vast amounts of data, and can optimize triage. In this context, AI applications are developed to train data-driven patterns in either unsupervised or supervised learning, utilizing past data via various machine learning paradigms such as deep learning, support vector machines, or recruitment of species-based algorithms; autonomously software predicts the outcomes, which can be used to categorize patients by urgency levels. AI technology is also leveraged for non-traditional triage scenarios such as prehospital identification, out-of-hours screening, and setting hundreds of miles away from EDs.

5.2. Impact on Emergency Care

Although not the origin of modern medicine, emergency care as we know it today is based on principles and practices that seek to mimic the system of triage first established by the military during battle or mass casualties [1,2]. States realize an actualization of the principle of socialization of health rights through the establishment of universal public health systems that guarantee all individuals who need it access to timely emergency care. Triage is, therefore, the frontline of these systems, receiving all demands on an equal footing and determining who will be treated and when, in this case by identifying patients who need urgent intervention. Health professionals at the triage station make decisions that can lead to two different results: there may or may

not be an associated risk of death within the following 24 hours, and the associated mortality of patients who are not operated on can be as high as 11%. The transparency implicit in the idea of socialization of health rights should lead to a minimization of errors due to poor patient stratification imposed by precedent “less urgent” cases that increase the waiting time for urgent patients. In this way, thanks to this and the monitoring systems, the health system assumes the risk of allocating human and economic resources in the event that patients arrive who do not generate a cost and the triage coordinator will validate the decisions taken at the time of treatment.

Errors during triage are possible due to the potential human error of the nurse, aided by the pressure of many people waiting with varying levels of social urgency, the experience and knowledge of the health professional in the field, and the high uncertainty during the initial moments of seeing the patient. Recently, Decision Support Systems have made their appearance in this zone of potential human error that justify and support the triage decision to reduce the number of associated mistakes. Emergency departments present a substantial variability in the arrival of patients and their demand for resources over time accompanied by an increase in demand and workload, so the post-COVID world has highlighted even more than before the need to establish Decision Support Systems. These have been developed in many fields of emergency medicine: clinical decision rules, severity scales, disposition decision support, clinical prediction rules, risk adjustment systems, trauma scoring and description systems, patient outcome prediction, prognosis models, risk stratifying for airway, ventilation or adrenocorticotropin hormone pump failure, and psychological support or psychiatric evaluation. Common all these systems is the development of mathematical models that are capable of processing massive amounts of data very quickly and accurately.

5.3. Ethical Considerations

As AI-CDSS control the most challenging and critical cases in emergency care, the consequences of errors and the harm they may cause is extreme, generating high ethical responsibility. AI-CDSS may drive false-negative errors, forgone resources, dropout communication, and poor quality of the emergency care processes. AI-CDSS may drive false-positive errors and overload emergency care, decreasing quality while augmenting the risks of unwanted effects in cases of children or elderly people, benign cases, and critical but rare cases. These problems are increased due to the variability of CTAS at some times of the day, driven by the ebb and flow of resource availability and not correspondingly adapted by the center’s activity. Some AI-CDSS suggestions may be wrong, and expert physicians assume the responsibility by balancing the algorithm priority with their own clinical experience. Is there a room for a lack of consequential

damages responsibility transfer for specific types of AI-CDSS? The question will have to be dealt with and often, policies are the only way.

The interactions between humans and AI-CDSSs bring up questions regarding direct and indirect decision responsibility, especially about 'recommendations'. AI-CDSSs sophistication allows a more complex type of recommendation based on the possibility of interactive learning. Moreover, as the knowledge of the MD becomes progressively aligned with the AI-CDSS over a period of increased repetitive practice, a behavioral nudging effect could arise that may be more powerful than simple ML mechanisms. Some enthusiasts defend that MDs should adjust their decisions towards those suggested by complex AI-CDSSs, assuming downward consequential damage responsibility. We must also question what it means to become a physician, and whether the reduced learning curve made possible by reinforcement learning nudging AI-CDSSs will bring diminishing returns after years of practice.

6. EHR Integration

6.1. Importance of Electronic Health Records

Clinical decision support systems can substantially aid decision making, but the input about a patient must be informative and up to date for the support to be clinically relevant. Data about a patient's medical history and current state should be easily accessible because without adequate information, decisions may be poorly informed and mistakes can occur. Electronic health record systems have become ubiquitous in the practice of medicine, appearing in nearly every clinical setting from primary care to specialized healthcare facilities. An EHR contains detailed demographic information as well as clinical data about a patient, including past medical history, medications, medical allergies, immunizations, laboratory test results, radiology images, as well as billing information. EHR systems have become a centralized depository for much of the typical clinical information needed by health care providers. Despite their centralization of clinical information, EHR systems differ significantly between healthcare systems and even among practices. These differences may complicate the integration of CDSS within EHRs, a necessary requirement to make their use simple for the physician user.

Computerized physician order entry modules are typically bundled within EHRs so that they can share data about a patient directly with ordering providers when needed. Many clinical decision support interventions are used in conjunction with CPOE, such as interaction-side alerts. While EHRs generally aid the integration of decision support interventions that display during the course of ordering, not all CDSS are located directly within the EHR. Examples of CDSS that are distinct from EHR systems include those that are used in an analysis of big data sets for predictive modeling of

adverse outcomes, as well as CDSS that are developed to provide clinical support only via a web interface without the ability to access the EHR directly. Such systems may be unlocked remotely into the web-based interface by providers with varying degrees of automated interfaces with EHRs, or not unlocked at all. Even in the setting of CPOE, the use by clinicians of CDSS that operate outside the EHR while not much slower than CDSS that work seamlessly with EHRs is less appealing.

6.1. Importance of Electronic Health Records

Electronic Health Records are digital files that catalog patients' medical history and status, storing information on demographics, medications, allergies, medical history, progress notes, immunizations, laboratory data, radiological images, and billing information, among others. The Clinical Decision Support System interacts with these data to enhance the patients' safety and quality while disseminating knowledge with a service posture, reducing variations in practice and implementing efficiently evidence-based guidelines. There are different types of CDSSs: those that provide a diagnosis suggestion, usually treating a specific disease; those that support therapy; those that suggest treatment changes, frequently in the contexts of intensive care units; those that provide reminders; and those that plan tests. In most cases, these decision support systems are fed by EHR.

Electronification of clinical records aims to ensure that information about patients be easily accessed whenever needed and from any possible device, since the use of medical records is governed by ethical and legal aspects. The strategy chosen to execute this electronification is the use of Electronic Health Records, a digital system that regulates the storage, maintenance, and sharing of patients' data. An EHR is defined as a digital record of a patient's health information that can be shared across different health care practices, including hospitals and medical clinics. Starting as a digital file of handwritten records, EHR have evolved to become a multi-feature tool that connects various health care providers and encompasses all medical data related to a patient.

6.2. Strategies for Effective Integration

A large number of researchers are involved in the integration of components into the existing systems. However, not many seem to be actively pursuing decisions regarding the content and the methods used within it. Decisions are rarely made regarding suitable selection of components or the appropriateness of the methods employed. Often the only focus is on the technical aspect of integration without regard to the functional aspects. To achieve the functional level or reach a similar status as the original systems, the external appearance and working processes need to be flexible as

briefed earlier. The system must return the required data in the appropriate manners for use by the clinicians. In addition, it must also allow the users to view it in a proper, coherent format without confusing errors.

The system can supplement both the clinicians and patients with the display of other preparation files such as flow charts, self-assessment jazzed up with motivational messages, literature references regarding the treatment protocols. These files cut down the consultant's time spent on the system and better the patient's experience while motivating discussions. Generation of multiple components seems unlikely-to-usability issues, long loading times, cluttering the system and distractions. Thus, whereas having a wide database may help improve accuracy, putting all the possibilities on display may only help worsen the usability. The ease with which a design can be understood is determined by the user's previous experience with similar designs. The goal would be to create a visually engaging product that provides users with quick access to needed information. While the generation of multiple output components seems unlikely-to-usability issues, long loading times, cluttering the system and distractions. Thus, whereas having a wide database may help improve accuracy, putting all the design permutations on display may only help worsen the usability. The ease with which a design can be understood is determined by the user's previous experience with similar designs.

6.3. Case Studies of Successful Integrations

The integration of PI and CDSS tools for diagnostics greatly improves the workflow quality and clinical aspect. The capability of performing decision support in the same place, at the same time, with the same patient context as a clinician decision, greatly increases the reach, acceptance, relevance, and indeed the number of decisions that could be supported.

A great example of successful integration of a clinical decision support tool with a widely adopted EMR platform is the diagnostic system developed for the Partners Longitudinal EHR dataset from 2000. In this case, the CDSS runs at the same time and in the same interface as the EMR point-of-care use. In this way, it perfectly overcomes all the integration problems previously cited above. Any patient data in EHR can be used for the diagnostic knowledge base and decision making. The support and reach for rare diseases achieving the final diagnosis may be lifesaving, decreasing the average time to diagnose is also clinically important. On the other hand, the clinical reuse of the diagnostic decision increases the opportunities of learning associated with that case to optimize the energy spent, reducing the proportion of undiagnosed patients, avoiding risks for the patients. The integration of CDSS tools with the usual clinical practice is crucial to success.

The experience of the old ISAhD system, and of successive similar attempts has been that the only real way to achieve integration of emergency diagnostic systems with EMR systems is to integrate the two together. The reason for this is that the decision of which services to run is not selectable by the clinician on clinical grounds; an emergency patient will not comply with qualitative criteria that would allow a proper prior selection of cases.

7. Multi-Modal Fusion in CDSS

Numerous amounts of costly, low signal ratio data make up the plethora of clinical data recorded by the electronic health records systems, hospital picture archiving and communication systems, digital monitoring systems, among others. The magnitude of clinical data, therefore, necessitates a thoughtful integration process. Multi-modal fusion describes a variety of data-centric approaches to effective information integration and synergy creation, allowing the optimized clinical guidance of decision support systems, and is a complex, vital layer of customized clinical decision support systems. Various centralized multi-modal fusion methods address different data formats, be they time series, digital voxel representation, test results, visualizations, and image and spatial data, combining them into fused digital pathophysiological models. Such data-driven pathophysiological models synthesize large clinical datasets into minable knowledge representations. Fusion functions can be any combinations of deep multi-task learning, sparse and multi-modal clustering, support vector machines, kernel models, graph searches, and subspace projections, and delineate standard function outputs and shared, domain-specific knowledge.

The classical aspect-oriented multi-modal integration of data synergistically contributes to the fusion of task-specific classifiers and pathophysiological representational aspects, able to effectively recognize and model clinical decision points such as diagnosis, progression, intervention, and outcome, reflected by the multitude of clinical events. Multi-modal clinical data fusion is facilitated by multi-speaker communications and clinical item pool discourse-based multi-modal clinical data integration methods. Synthesizing individual-factors and representation interface layer aspects allows the domain specialization of clinical decision support systems and fuses user-friendliness and representational-interpretable, user-involvement template assets into inception. Data fusion allows input-distribution budging capabilities, obscuring deficiencies in individual input classification sufficient for clinical task performance and user-involvement capabilities.

7.1. Definition and Techniques

Fusion is a general technique used to combine data from limited sources and draw a conclusion that is more informative and complete than that provided by any one of the

individual sources. Multi-modal data fusion in Clinical Decision Support System (CDSS) refers to the integration of observations made through the different modalities of decision support in healthcare to achieve better inference performance. This usually involves the mapping of heterogeneous high-dimensional data defined on different spaces to the seminal low-dimensional semantic space. The sources of information could be sound, vibration, pressure, optomechanical, electromagnetic wave, electrochemical, electromagnetic radiation, electromagnetic charged particle, and visual. The uncertainty that is inherent in these observations may stem from a myriad of reasons.

Multi-model fusion has become a peculiar application-specific challenge of considerable practical importance and it continues to receive significant attention from researchers. It employs a wide variety of automatic and semi-automatic techniques and strategies, depending on factors such as availability of sufficient labeled data, the reliability of the individual sources, the diversity and quality of the classifiers being combined, computational resources, and the specific application. The most common techniques include majority voting, linear regression, and decision template or distance-weighted learning framework, and neural networks, among others. Multi-model fusion, although a relatively simple tactic, has proven to be surprisingly effective in practice. It is routinely employed to combine the predictions of these classifiers for solving specific applications in biology and medicine, including tissue classification from gene expression data and imaging analysis.

7.2. Benefits of Multi-Modal Fusion

AI-driven clinical decision support systems (CDSS) generally tackle a single task, typically based on one specific type of data or information input. This is also true for the majority of the existing AI applications using clinical data, employing only EHR for clinical diagnosis prediction, or using only medical imaging for malignancy detection or language models for trial recruitment. Most of these systems perform accurately at best for some specific tasks in domains where large-scale labeled data are available. However, for some domains, large-scale predictions are not feasible, and for some clinical prediction problems, classical single-modality task predictors are not independently reliable and may not play significant roles. When predicting a task that requires different types of clinical data or information input, or when conducting an uncertain task, CDSS would definitely benefit from the collaborative contribution across other types of inputs.

In fact, multi-modal multitask prediction has been a far right trend across other non-clinical research areas, especially in the vision and language fields where tasks such as image caption generation, visual question answering, and visual grounding have

proved their effectiveness of multi-modal multitask learning and prediction, providing complementary supervisory signals from disparate modalities. Typically, prediction tasks in different domains are inherently linked. In the clinical area, there exist many naturally associated tasks such as those involved in a clinical pathway, e.g., disease diagnosis, treatment and progress detection, patient profile description, clinical trial recruitment and outcome prognosis. Such tasks are closely related with each other, requiring disparate, yet interleaved clinical data.

7.3. Examples in Clinical Practice

The medical domain is complex and very tightly constrained, both for the prediction problems as well as for the data modeling. Multi-modal models in the medical domain typically make use of vision and text or tabular data inputs together, since these modalities often provide complementary information, even in the same patient visit. Many existing works address two of the three branches, mostly because of the limited availability of appropriately pre-processed large-scale datasets.

In the study, CT image data of the abdomen and head are used with EHR data of clinical inquiries in a critical cohort to predict three frequent medical inquiries within one week after the CT scan, namely a lab exam and two different possible therapies. The experiments showed that contrast-enhanced CTs together with prediction for the clinical inquiries considerably improve the prediction quality compared to prediction based on the relevant EHRs alone. The methods proposed suggest a ranking approach to suggest which clinical inquiries need to be addressed, since different patients require rather different medical attention. This work describes a macro-level approach to dropdown menus which can be supplemented with expansions at a micro level.

A method was devised that considers to fuse EHR and image data but without meta-information on the images included in the training sets when training the models, in order to predict ICD codes of the relevant time periods from the EHRs alone. In contrast, we consider meta-information provided by the EHR data together with image data in order to predict patient inquiries. The fusion models proposed consider multi-modal data to build a multi-modal representation, while the models estimated in this work aim to modulate the predictive quality by jointly considering the EHRs and ground truth inquiries as auxiliary non-image inputs.

8. Evaluating Impact on Patient Outcomes

The majority of CDSSs have been externally validated to check whether they are usable at scale, however assessing their impact on patient outcomes is more difficult and is not done as frequently. This is arguably the most important question researchers wish to answer since if an AI-based CDSS does not help physicians make better

decisions, it is not worth deploying at scale. There is ongoing discussion in the AI for Health community about how best to implement evaluations and what types of evaluations are useful. Many high-profile ML health research studies evaluate the performance of the algorithm using metrics out of sample. These are only one of many steps and are not enough in themselves. Such test set evaluations assume that the first thing is the only thing that matters. However, while ML algorithms have been created that are very good at prediction tasks, it does not follow that such algorithms will lead to a measurable improvement in health outcomes when deployed at scale, for the first thing is not the only thing. Health outcomes are affected not when an ML system merely answers a clinical question correctly, but rather through their effect on downstream clinical actions. These in turn result in changes to clinical pathways and, ultimately, the health outcome of patients. This suggests that understanding the quantitative relationship between predictive performance and patient outcomes is crucial for deploying clinically useful ML algorithms. A high-performing algorithm may be useless if it is not used to change clinical actions, as missing the first step may be the only achievable outcome. Out-of-sample accuracy is a poor proxy for impact, as an accurate model may be used to suggest interventions that are not actually helpful in the real clinical pathway.

8.1. Metrics for Evaluation

We have discussed evidence of effects of CDSS on test ordering, adherence to prevention guidelines, prescription of indicated therapies, and disease diagnoses. Here, we discuss the relationship of CDSS to patient clinical outcomes. CDSS were initially introduced as a response to the high level of errors in medicine. Since then, advances in software techniques for machine learning and improvements in the efficiency of certain diagnostic tests have reduced this error rate. Given these changes, we explore the continued value of AI-Driven CDSS in improving patient outcomes.

The use of CDSS has been associated with improved patient outcomes such as reduced mortality and morbidity, as well as fewer hospitalizations and readmissions, but such studies are few in number and often unsatisfactorily controlled. One reason for the low number of studies may be the difficulty of conducting them. Measuring medical errors requires costly and intensive studies that are likely to eliminate the resources that might come from the given organization as a result of the reduction. But the resources needed to study effects of CDSS on mortality and spending are also sizeable. We discuss the effects of different metrics and measures, as well as study designs, on our weight. CDSS research is focused on higher level outcomes, such as mortality and readmissions. These other outcomes are also determinants of spending and CDSS use, but we seek to determine their impact on patient risk factors and other determinants of spending.

8.2. Longitudinal Studies and Findings

Many studies and projects are aimed at documenting the clinical, organizational and workflow changes that occur over time, impacting patient outcomes after the deployment of a CDSS. These longitudinal studies may take form of pre- and post-deployment comparative chart reviews, observational studies or controlled trials. These studies allow for the measurement of multiple types of patient and system workflow changes across time and space and at clinical and organizational scales, providing the user with detailed how, when, where and why answers. In addition to a multitude of different CDSS and evaluation criteria, varying time scales and local, cultural, organizational and infrastructure differences are likely to contribute to the range of the reported positive, neutral and negative impacts of CDSSs.

A number of studies have been published that discuss and report on the outcomes observed following the deployment of the most common commercially available CDSSs. In this section, we discuss both more specific longitudinal studies as well as notable outcomes studies. Although there is a real need for additional replicative as well as negative studies reporting on the impact of CDSS deployment, introduction and monitoring of the most common commercially available CDSSs in the routine decision-making process, the ability to longitudinally assess the impact of a CDSS on healthcare practices is becoming increasingly more difficult. This reflects both the dramatic changes in how healthcare is delivered and compensated for, the rapid pace of introduction of new systems used in clinical workflows and the continuously accelerating rate of in-processing of medical data. As such, CDSS monitoring for impact is often reduced to reporting surrogates found embedded in regional or national health databases.

8.3. Barriers to Effective Evaluation

Evaluating the impact of AI-driven clinical decision support systems (CDSS) on patient outcomes can be challenging. While evaluating CDSS evaluation on technical components and performance metrics like calibration and discrimination are common, these metrics fail to show the real-world effectiveness of algorithms when integrated with health systems and workflows, on the desired end-user group. No single method of evaluating AI-driven CDSS is sufficient; rather a phased-hybrid approach using qualitative methods to promote stakeholder engagement, quantitative studies based on simulation and/or retrospective data, and randomized and pragmatic trials, is needed based on resources and context. Furthermore, the healthcare stakeholders should have clearly defined a priori AI-related goals and challenges, as well as clinical domain expertise, when validating and evaluating AI-driven CDSS, which is not routinely the case today.

Multiple other barriers exist that limit the ability to rigorously evaluate the impact of AI-driven CDSS on patient outcomes in clinical settings. The lack of appropriate guidelines for evaluation, rapid technology prototyping and testing, and complex human-AI and system interactions can distort measurements associated with AI CDSS. There are also intrinsic difficulties in studying real world, complex adaptive systems with rigorous experimental designs, in real world settings. The physicians' limited and skewed availabilities during the clinical workday makes it daunting to conduct rigorous experimental studies. It can also be difficult to ascertain the true impact of AI CDSS on human behavior when it is just one of the many competing influences. Differences in patient races and histories, and physician selection, availability and experience, can also lead to selection biases in randomized controlled trials.

9. Future Directions in CDSS

Advances in technology and methodologies for managing data as well as for machine learning and AI present new opportunities for CDSS to contribute to clinical care in important ways. Thus far, CDSS that are heavily integrated into the clinical workflow are still quite simple. Moreover, CDSS in general have relied primarily on heuristic rules and regressions for the decision support they provide. Future generation CDSS may more directly leverage additional exciting AI platforms utilizing predictive modeling, NLP, computer vision, and multimodal modeling of complex real-world signals as input. These capabilities applied to data at scale have been used to deliver state-of-the-art AI algorithms in fields such as natural language processing. The potential for such techniques to enable CDSS that leverage predictive modeling, NLP, computer vision, simulated subjective signals of clinical reasoning, and multimodal modeling for complex temporal signals has not been fully tapped. Healthcare is complex because it involves people and their behavior, emotions, and motivations. It is a work-in-progress that includes myriad data points from the past, present, and future. The challenges are significant, but so is the potential.

Equally important is the future potential of CDSS to help with personalized medicine. Clinical decision support has had an emphasis on guidelines; however, these can be derived mostly using averages and do not optimize care for specific patients. Instead, with advanced machine learning models using profiling of individuals drawing on deep phenotyping and precision medicine, clinical decision support can help optimize outcomes for individuals instead of the population at large. The challenges to this direction are the fact that precision healthcare often requires complex data-gathering and lengthy time-associated profiling and other data related to decision support need to be closely aligned with the timing of decisions by clinicians. Moreover, education and acceptance of the clinical community are key challenges, which extend to the future direction outlined before with respect to advanced AI-powered models.

9.1. Emerging Technologies

Artificial intelligence (AI) is an area of computer science research that includes such sub-disciplines as Machine Learning, Natural Language Processing, pattern recognition, knowledge representation, automated reasoning and robotics, among others. The goal of AI is to design and implement intelligent software systems that can accomplish tasks that are commonly associated with human intelligence. AI is increasingly being applied in diverse domains, and is now being utilized in many clinical healthcare systems as well. AI has made profound inroads into clinical decision support systems, which play a pivotal role in supporting clinician decision-making. There now exists an entire sub-field of research in AI specifically geared toward supporting healthcare practice, and it is known as AI for healthcare.

In this paper, we focus on clinical decision support systems, which are AI-powered tools that leverage on patient data to assist clinicians with their decisions. Such decisions include whether to order a certain test, or to treat a patient in a certain way, or what diagnosis a patient may have based on their symptoms, or what treatment plan would be most effective. The feedback from the clinical decision support systems may either be either an independent decision or a recommendation as an adjunct to the clinician's primary decision. While AI-based clinical decision support systems have been gaining traction, they are not flawless, and several of them have manifested unexpected failures, causing patient deaths and near-death situations in some cases. Despite these setbacks, we are in the midst of an AI-based renaissance in clinical decision support systems. These AI-clinical decision support systems have started to out-perform healthcare professionals in certain tasks, and the large-scale deployment of electronic health records is offering a wealth of freely available patient data on which to train and test these systems.

9.2. Potential for Personalized Medicine

Among the many potential uses for AI in medicine, development support tools could come to play a major role in the drive toward precision health and personalized medicine. Research on alignment of the gut microbiome with dementia risks, genetic testing used to individualize treatment for cystic fibrosis or for hereditary breast-cancer syndromes, use of genotyping to minimize risk of malignancy for patients being treated for lymphoid malignancies, or treatment recommendations for cardiovascular disease based on polygenic risk scores are only a handful of the current areas of research in medicine that support such hopes. Numerous other disease areas are also being funded, including use of genetic testing for individual risk or treatment of sickle-cell disease, genetic testing to predict treatment efficacy of oral anti-coagulants in atrial fibrillation, or for risk of severe sepsis. Potential developments in pharmacology and

genomics, such as the emergence of drugs that selectively inhibit a specific mutation in a precision-treated population of cancer patients rather than a bulk cytotoxic agent for all patients, may increase the need for CDSS in formalizing implementation of genetic testing.

However, as with other new disciplines, the pace of expectation from AI-driven CDSS development growth and extrapolation has already exceeded its knowledge base. Summary gradient boosting has only recently brought forward any discussion of machine learning methodologies in the guidelines to standards of care for children with obesity. Identification of matching genetic targets in patients with cystic fibrosis, or approval of the HER-2 containing neoadjuvant chemotherapy regimen, has only recently entered into national standards of care. The statement that “targeting, not typing, is the goal of the future” for development of precision drug recommendations underscores the fact that recommendation for a specific mutation or molecular pairing is not the replacement for comprehensive precision-guided treatment for every cancer patient when the target is present.

9.3. Regulatory and Policy Considerations

As the use of AI and ML in healthcare expands, the regulatory environment will need to keep pace with the rapid speed of development in research and innovation. The regulatory environment for CDSS is a complex interplay between regulation of software as a medical device and clinical evaluation of premarket testing. Related laws govern the scope of product reviews and authority to regulate. There are draft guidance documents with policy recommendations, examples, and risk categorizations. It is clarified that only devices and software intended for medical purposes that are within the definition of medical devices are regulated. These devices must have a "therapeutic" or "diagnostic" intent. However, the drafts imply that there is no intention to regulate devices that are not intended to be used to help a healthcare provider make a clinical decision about the diagnosis or treatment of a given patient.

While there is authority to regulate AI-enhanced CDSS, there are limited resources and authority. Like the regulatory body, other organizations are also interested in oversight of AI-enhanced CDSS. One manages reimbursement and describes the need for high-quality evidence. Another plays a coordinating role with respect to the health IT ecosystem and safety. Many observers believe that further government intervention into the AI-enabled CDSS market could harm the development of the technology. Other experts believe it essential for several reasons, including infrastructure development and accelerated development around population health, quality, and lower costs.

10. Conclusion

The rapid growth of societal need for healthcare indicates that despite the advances in technology and understanding of disease, the deliverables of national biomedical science programs need synchronization with approved economic models and frameworks to generate evidence for their adoption as part of fundamental clinical practice. Clinical decision support systems could thus facilitate augmenting current clinician workflows and enable a greater success of newer and novel diagnostic and therapeutic options while ensuring clinical accuracy and non-maleficence for the most vulnerable. A major step towards such directions was taken recently during COVID where immense and rapid collaboration borne out of necessity between traditional academic institutions, tech companies, and regulatory agencies resulted in unprecedented moves from development to approval and real-world implementation. These new tools now have the potential to aid clinical practice in novel ways including those in areas like diagnosis, procedures, hospital workflows, patient risk, and help in remote patient monitoring tools for long hauler patients.

Despite the omnipresence of AI and complex algorithms in our everyday life, the understanding of AI-driven CDSS is still in infancy. There lies an immense opportunity for these new augmentation technology platforms to drive adoption and success of new preventive, diagnostic, and therapeutic options while ensuring that the promise of better healthcare systems is delivered. In this summary, we strive to demystify these tools in their evidence-based fundamentals for use in clinical practice and provide an overview from both a high level as well as easily understandable summaries by area for the busy clinician. We finish with future perspectives as this supplemental technology is poised to rapidly influence all aspects of patient care.

References

[1] Ouanes K, Farhah N. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. Journal of medical systems. 2024 Aug 12;48(1):74.

[2] Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, Gilbert TK, Hagendorff T, Holm S, Livne M, Spezzatti A. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health. 2022 Feb 17;1(2):e0000016.

[3] Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Frontiers in digital health. 2021 Jun 25;3:645232.

[4] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X. Reporting guideline for the early stage clinical

evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. bmj. 2022 May 18;377.
[5] Haick H, Tang N. Artificial intelligence in medical sensors for clinical decisions. ACS nano. 2021 Feb 23;15(3):3557-67.
[6] Schwalbe N, Wahl B. Artificial intelligence and the future of global health. The Lancet. 2020 May 16;395(10236):1579-86.
[7] Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. The American journal of medicine. 2019 Jul 1;132(7):795-801.
[8] Yang Y, Siau K, Xie W, Sun Y. Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. Journal of Organizational and End User Computing (JOEUC). 2022 Jan 1;34(1):1-4.
[9] Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, Baeshen HA, Sarode SS. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making-a systematic review. Journal of dental sciences. 2021 Jan 1;16(1):482-92.
[10] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. Journal of global health. 2018 Oct 21;8(2):020303.
[11] Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019 May 1;1(1):e13-4.
[12] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017 Dec 1;2(4).
[13] Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of medical Internet research. 2020 Jul 29;22(7):e18228.
[14] Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020 Nov 2;35(42).
[15] Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020 Feb 25;98(4):251.
[16] Bohr A, Memarzadeh K, editors. Artificial intelligence in healthcare. Academic Press; 2020 Jun 21.
[17] Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. Jama. 2020 Feb 11;323(6):509-10.
[18] Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ digital medicine. 2018 Oct 2;1(1):53.
[19] Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019 Jul 1;8(7):2328-31.
[20] Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, Malhotra N, Lui V, Gibson J. Artificial intelligence for good health: a scoping review of the ethics literature. BMC medical ethics. 2021 Feb 15;22(1):14.
[21] Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype?. Jama. 2019 Jun 18;321(23):2281-2.

[22] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun 1;6(2):94-8.
[23] Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020 Nov 30;20(1):310.
[24] Hosny A, Aerts HJ. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955-6.
[25] Ho A. Are we ready for artificial intelligence health monitoring in elder care?. BMC geriatrics. 2020 Sep 21;20(1):358.
[26] Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British medical bulletin. 2021 Sep;139(1):4-15.
[27] Lau AY, Staccini P. Artificial intelligence in health: new opportunities, challenges, and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):174-8.
[28] Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:100099.
[29] Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, Mccradden MD, Oakden-Rayner L, Pfohl SR, Ghassemi M, Mckay F, Treanor D. The value of standards for health datasets in artificial intelligence-based applications. Nature medicine. 2023 Nov;29(11):2929-38.
[30] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. InHealthcare management forum 2020 Jan (Vol. 33, No. 1, pp. 10-18). Sage CA: Los Angeles, CA: Sage Publications.
[31] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
[32] Sahni NR, Carrus B. Artificial intelligence in US health care delivery. New England Journal of Medicine. 2023 Jul 27;389(4):348-58.
[33] Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics. 2019 Feb 1;21(2):121-4.
[34] Benke K, Benke G. Artificial intelligence and big data in public health. International journal of environmental research and public health. 2018 Dec;15(12):2796.
[35] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization. 2020 Jan 27;98(4):257.
[36] Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health. 2021 Jan;18(1):271.
[37] Shivadekar S, Halem M, Yeah Y, Vibhute S. Edge AI cosmos blockchain distributed network for precise ablh detection. Multimedia tools and applications. 2024 Aug;83(27):69083-109.
[38] Panda SP. Augmented and Virtual Reality in Intelligent Systems. Available at SSRN. 2021 Apr 16.

[39] Panda SP. The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR). 2025 Jan 1.
[40] Shivadekar S, Kataria DB, Hundekar S, Wanjale K, Balpande VP, Suryawanshi R. Deep learning based image classification of lungs radiography for detecting covid-19 using a deep cnn and resnet 50. International Journal of Intelligent Systems and Applications in Engineering. 2023;11:241-50.
[41] Rane J, Chaudhari RA, Rane NL. Data Analysis and Information Processing Frameworks for Ethical Artificial Intelligence Implementation: Machine-Learning Algorithm Validation in Clinical Research Settings. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:192.
[42] Mohapatra PS. Artificial Intelligence and Machine Learning for Test Engineers: Concepts in Software Quality Assurance. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:17.
[43] Nuka ST. Next-Frontier Medical Devices and Embedded Systems: Harnessing Biomedical Engineering, Artificial Intelligence, and Cloud-Powered Big Data Analytics for Smarter Healthcare Solutions. Deep Science Publishing; 2025 Jun 6.
[44] Maguluri KK. Machine learning algorithms in personalized treatment planning. How Artificial Intelligence is Transforming Healthcare IT: Applications in Diagnostics, Treatment Planning, and Patient Monitoring. 2025 Jan 10:33.
[45] Panda S. Observability in DevOps: Integrating AWS X-Ray, CloudWatch, and Open Telemetry. International Journal of Computer Application. 2025 Jan 1.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 7: Artificial Intelligence Models for Meteorological Data and Climate Patterns

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Meteorological Data and AI

This section provides an introduction to meteorological data variables, weather modelling, and AI artificial system applied to meteorological data. For weather forecasting to technical design, meteorological data provides critical information describing atmospheric conditions probably affecting certain locations during a specific time period. They may be provided by means of signals addressing any of the fundamental atmospheric attributes, including temperature, pressure, humidity, wind speed and direction, precipitation, horizontal or visibility, and clouds. The applied modelling AI system addresses three model types: detection or monitoring, now-casting and forecasting.

Earth surface weather is one of the most pronounced manifestations of the state of the planetary atmosphere. Weather clouds or determines the success of crops and animal breeding technologies crucial for human sustenance [1-2]. Weather reveals the state and behavior of the atmospheric substance enclosing the planet and creating the forces of the Sun revolution and Moon swirl needed to maintain its orbital dynamics. Weather shapes the spatial climate structure and determines climate change in time, as well as the ecological stability of terrestrial varieties of the living world that took form through life evolution.

It has long been understood that weather implies continuous short-term variations of local and global ecogeophysical parameters around their long-term average which are averaged as climate [3-5]. However, weather, which is established through the interactions of the Sun and Moon with the atmosphere and Earth, equally shapes the

underlying physical processes governing climate active diagnostic indicators, such as temperature, pressure, humidity, wind, precipitation, and the reflection of solar radiation by clouds or the surface beneath. Despite the fact that artificial intelligence may provide improved solutions for some or most aspects of meteorological data modelling, solutions are impossible without clear understanding of the way satellite data is used for solving these problems.

2. Spatiotemporal Deep Learning

Deep learning models have had great success in many fields of research, and new architectures appear frequently. Classical CNNs work on individual images, holding a lot of spatial information [6-8]. However, in applications such as weather forecasting, satellite image analysis, or air pollution analysis, we have image sequences where the temporal patterns improve the understanding of the spatial patterns and vice versa. Several works showed that including spatiotemporal modeling improves the results of 2D CNNs for satellite image analysis. This lines of research open the field to any CNN temporal extension. In this section, we briefly explain some of the recent temporal models aimed at scientific tasks that are based on ConvLSTM and attention mechanisms.

The Convolutional Long Short-Term Memory Networks (ConvLSTM) is an extension of the LSTM. Both models represent streams of information. Considering that most of the scientific problems for which spatiotemporal modeling is required hold data in image form, the authors included 2D convolutions in the LSTM gates to preserve the spatial information. Since then, several success stories of the ConvLSTM architecture in scientific applications have emerged. Even when variant models are sometimes more convenient, due to the success and simplicity of the design, ConvLSTM-based models are usually the first option, and have become the reference spatiotemporal models. Therefore, the architecture has gone through several deep learning innovations [7,9-10]. A few works reduced the number of internal LSTM cells due to the number of parameters of the 3D CNN. Furthermore, some studies proposed a different fusion mechanism, or discarded the LSTM cell altogether with no significant drop in performance. Other variants proposed using 3D convolutions for spatiotemporal encoding. Finally, some methods proposed adding ConvLSTM to a backbone model or made it conditional to generate better results. Nevertheless, most of these works still use the original model or variants closely related to it.

2.1. Overview of ConvLSTM

Capturing spatial and temporal dependencies of data is critical in different areas of research. In many cases, the data present spatial dependencies because of multiple spatial locations and sensors. Gathering information from many sensors, present at

different locations creating data set, allows the creation of a data set containing temporal dependencies; a data set collected of objects that saluted at different times, creates a data set with temporal dependencies. The manipulation of such data set is rich and allows better solutions to real world problems [1,11-14]. For example, across the forecasting science, the study of spatiotemporal dependencies is useful for predicting weather, where the meteorological conditions of each point in space are related, and change across time. Predicting the demand of certain services at a certain point in time is currently used by delivery applications to supply those services; supplies in multiple points can be predicted simultaneously, by creating a data set with multiple just-in-time supplies created simultaneously. Many approaches have been created that allow address spatiotemporal dependencies created by data sets containing spatial and temporal data with dependencies. Therefore, understanding how to model these data sets is important.

To understand how to model this type of data sets is why methods have been proposed; first was the convolutional neural network temporal group, those proposed just to create temporal group encoders with limitations for video analysis, these limitations moved the deep learning community to propose the three dimension convolutional neural network which can be used for video classification; both approaches only deal with spatial and temporal dimensions separately. In both cases, the loss function calculates the score for temporal dependencies globally or for short sequences; actually making them not able to represent the true temporal dependencies of the data. Next, the ConvLSTM model was proposed; the approach was designed to be used for specific applications where it is required the prediction of radar image sequences; aiming for the inclusion of spatial and temporal dependencies to improve current predictions. The model was designed for memory purposes; ConvLSTM allows the setting of data specific to the spatial group, controlling how many dependencies to attend separately on each of the dimensions. ConvLSTM aimed to be easily applicable with a good performance in prediction; to achieve good performance, it applies LSTM loops to 2D convolutional layers allowing any type of application in spatiotemporal forecasting.

2.2. Attention-Based Models

The attention mechanism allows a model to dynamically compute temporal and spatial weights while processing an input with strong correlations. Attention-based models outperform ConvLSTM in capturing longer spatial-temporal relations. Modifications to the traditional cross attention mechanism in computer vision create a new spatiotemporal attention block [13,15-17]. Fusing the attention module with transformers, the model is applied to image prediction tasks and shows significant improvements in long-term temporal relations over previous methods.

2-D attention is fusing satellite and weather data to predict 1-h precipitation. The model integrates the pixel-level correlation learning efforts of the convolution layers with attention modules to establish both temporal and channel relations [18-20]. Factorized spatiotemporal attention is used to predict precipitation, where the spatial attention weighs each modified 2-D observation for every forecast moment, and temporal attention concentrates on the immediate previous forecast, while fusing the weather input for each time step. The model automatically learns the multiscale relation properties. These studies show that attention-based models are promising approaches to contemporary meteorological forecasting.

The different capability of transformer in capturing various correlation properties drives abundant following research. Knowledge-based spatiotemporal transformer with explicit temporal and spatial knowledge about covariates is learning towards benefiting application. The proposed model integrates the capabilities of four transformer variants, with temporal, spatial, and the modified channel correlations embodied. The proposed transformer with the channel-wise attention weighted by additional knowledge weights outperforms previous models in daily precipitation prediction. A transformer-based model for spatiotemporal weather prediction exhibits lower particle count, lower Euclidean equivalent radius error, and lower normalized intensity error when the forecast time increases when predicting the simulation in the meteorological visualization field.

2.3. Applications in Meteorology

Various types of meteorological data have been the target of a long term consideration in both temporal and spatiotemporal research bottlenecks. Weather has a complex spatial structure due to presence of boundaries including land, sea, mountains, and also the time series that curb the ability to apply any discriminative method on raw data in a pre-trained, domain specific fashion, while which in theory can be used in any weather conditions, both supervised and unsupervised learning for any of these tasks. Although spatiotemporal attention models and ConvLSTM Neural Networks have been sparingly used on the initial tasks mentioned, such as clouds and storm prediction and precipitation accumulation prediction, Meteormatics is one of the few which directly deals with precipitation and cloud prediction.

To the best of our knowledge, the first applications of spatiotemporal models on numerical weather forecasting was based on a simple Seq2Seq architecture using a 2D CNN for both encoder and decoder, directly producing weather satellite images. They applied the model on a single forecasting point in time, and during the evaluation phase, they produce forecasting at several stages in the lead-time of the forecasting

task. In a related publication about satellite forecasting, some specific architecture tweaks are introduced [19,21-22].

Spatiotemporal stack of Neural Networks is used, where the study mainly consists of evaluation on rescaling monumental weather input and output datasets, by loss quality and quantification of the direct downscaling task from global scale maps to countries' level maps. On evaluation, they show the effectiveness of neural networks for the task, when dealing with DenseNet compared with traditional methods, yet, the task is still trivial with anything pretrained for the task.

3. Extreme Event Prediction

Although AI models can be used to draw a link between any two climate variables, there are specific areas of application for which AI models are better suited than traditional statistical models. These are areas where a traditional model is seen to struggle. For example, when using statistical correlation to study climate patterns, there is no means of discerning whether a data series for one variable is correlated with a different data series of the other variable, or whether both data series are correlated with a third, common external data series [11,23-25]. Also, AI models have been found useful for the making of extreme predictions when predicting climate data. Extreme predictions are used for making predictions of extreme values of a variable, such as annual maximum rainfall for a monsoon season. Some examples of extreme predictions for weather-related climate variables are monsoon onset and withdrawal date predictions, prediction of annual maximum rainfall, prediction of number of rainy days in a monsoon season, and tropical cyclone land-fall prediction. Monsoon onset and withdrawal date predictions have been carried out with great success using artificial neural network models, for locations along the southwest coast of India. In one study, we used an ensemble of relevance vector models to predict the monsoon withdrawal date at five locations along the southwest coast of India, using data of the long-range forecast model developed by the Indian Meteorological Department, and found that the importance of predictors was very different across locations. For Indian monsoon onset predictions, we have also successfully used sparse membership Gaussian mixture models. As stated earlier, for all the predictions mentioned above, statistical relationships have been found to hold good for a number of years, but not for the entire historical period. However, weather and climate are very sensitive to external inputs [26-28]. Climate model predictions change quite drastically if input external forcing, such as greenhouse gas emissions, is changed by a small percentage. Predictions of tropical cyclone formation and post-landfall intensities are very sensitive to variations in sea surface temperature, differential departure of middle tropospheric humidity from its mean value, and the mid-level temperature lapse rate.

3.1. Flood Prediction Models

Flooding represents one of nature's most devastating and unpredictable disasters, leading to immense economic losses as well as loss of human life. Frequently, satellite or radar images are utilized to detect gouges or changes in land cover, which can direct response teams in the field to ensure safety and minimize recovery time. In times of flood, visible and infrared bands detect variations in the surface due to stagnant water bodies [29-32]. The normalized difference water index (NDWI) shows the presence of water on the earth's surface by reducing the effect of infrared bands that are reflected off of surface features, such as vegetation. The NDWI map can identify flooded areas and track flooding propagation. NDWI maps might aid in identifying and mapping flooded areas or detect differences in specific areas by creating NDWI time-series, assisting post-disaster assessment.

Recently, concern has been raised regarding the lack of accuracy and lead time in operational flood forecasting. This lack of appropriate quality, location accuracy, or increased uncertainty has been recognized as important issues affecting the mapping of flood inundation areas. Therefore, the improvement of operational river gauge predictions is necessary before and during floods to ensure life and asset safety as well as the environmental protection for flood-prone areas. This research reviews recent advances in the application of artificial intelligence to flood forecasting and prediction in non-conventional approaches. This review is divided into three sections. The first section gives an overview of the use of AI technologies in flood hydraulics modeling with streamflow and water level prediction at ungauged sites and bridge and culvert performance [31,33-35]. The second part summarizes the application of AI in flood simulation and forecasting, from downscaling datasets to different temporal and spatial resolutions. The last section concludes with a review of future trends in the self-adaptive application of neural networks and deep learning in combined hydrology-hydraulic modeling, real-time forecasting with assimilation of remote sensing and satellite data, and forensic flood hydrology.

3.2. Hurricane Forecasting Techniques

More than a thousand hurricanes thundered through the tropical oceans of the globe during the past century, nearly 900 hurricanes earned enough strength to trigger destruction, and about 270 large hurricanes made landfall, causing horrific damage and mass casualties in the region. Understanding intricate hurricane mechanisms is a huge challenge because of the extreme nature of the events. AI techniques like deep neural models, statistical analysis, machine learning, and multi-factor risk assessments notably develop an outstanding effect in hurricane patterning, tracking, and forecast. A

combined type of different AI hurricane models used for different tracking and forecasting levels have proven to work highly efficiently and effectively.

Typhoon, hurricane, or tropical cyclone tracking and forecasting is a challenging and highly important task in meteorology. Poor prediction and underway monitoring can cause the loss of lives and extreme socioeconomic damage. Due to bulks of data available from brief historical events over the last century, hurricanes have been predicted using statistical and dynamic models currently in use. Conclusion prediction based on limited available hurricane data has a big risk of producing poor results. Since decades back, forecasting and tracking difficulty has risen with messy observations, longer hurricane seasons, the increase in number of tropical waves in association with heatings, weak upper-level winds, and climate variability observed. Deep learning models are being gradually adopted and highly valued due to their backup support for attention ensemble and hybrid models used in operational hurricane suppression planning and risk communication. The models have created an outstanding forecast effect in hurricane prediction skill. With quick technologic advancements, developments of more accurate satellite, radar, and buoy observing configurations and system network models are anticipated.

A key advance that made hurricane forecasting more reliable since this is the use of Advanced Dvorak Technique and Consensus Tropical Dvorak Technique that combine refinement in Dvorak techniques [36-38]. It was designed to help definitively identify and assess three-dimensional structure asymmetries in new-born tropical storms and weak, poorly-organized tropical cyclones with faint spiral banding and cloud tops. Adaption of Advanced Dvorak Technique and hybrid Consensus Tropical Dvorak Technique not only improves infrared and microwave pattern recognition but has strong ability to enhance feature-based infrared hurricane forecasting.

3.3. Wildfire Risk Assessment

Wildfires threaten lives, property, and ecosystem functionality across the world. While drought has historically expedited wildfires, other factors, such as higher temperatures, changing seasonal weather patterns, less rainfall, variability of wind patterns, and build-up of fuel favoring ignition and spread of wildfires are also contributing increasing damage from wildfires. The devastating effects of record-breaking wildfires have pushed governments, fire departments, universities, and community groups to focus on better containing and predicting wildfires. These efforts are not just limited to developed nations but are also prevalent in the developing world.

Data-driven wildfire initiatives have made it possible to undertake predictive modeling for initiation of wildfires and the employed methodologies such as GIS-spatial, data-driven, and simulation models focus on enhancing a predictive understanding of

wildfires [1,39-41]. Wildfire-preventive action models have focused on geospatial patterns of economic losses and climatic factor-mediated burn patterns of multi-hazard risk. With the global focus gradually shifting towards combating climate change, more and more efforts are being directed towards enabling sustainable measures that can be employed in limiting wildfires. In this spirit, researchers used deep learning methods to make predictions about climatic variables responsible for groundbreaking wildfires. Policy guidelines then employed the predicted climatic parameters to propose wildfire-control measures in the near future. In addition to predictive modeling, providing a risk assessment of wildfires based on observed climate features has been potent, in targeting and forewarning the decision-makers and helping with informed decision-making regarding wildlife conservation.

4. Ensemble Learning Approaches

4.1. Introduction to Ensemble Learning Data-driven methods for meteorological predictions usually rely on a single model - a myriad of model implementations of types ranging from simple statistical models to deeply layered artificial neural networks are used to forecast future meteorological behavior based on previously observed conditions. Ensemble Learning methods escape such single-model limitations, with the added benefits of improved prediction performance and more robust prediction uncertainty quantification. Ensemble Learning methods leverage information from multiple models to enhance overall performance and reliability. They do so on the premise that model-level diversity improves accuracy and reliability by reducing prediction variance and error [42-45]. There have been a few Ensemble Learning approaches applied to meteorological prediction in the past, but their applications have not led to outsized performances when applied in a simple manner. But as meteorological prediction gets increasingly more complex - higher dimensional spaces associated with weather variables, larger volumes of data collected through time, coarser temporal resolution - the application of Ensemble Learning in a proper and judicious manner, taking care to improve model diversity, and balance bias and variance of the prediction errors, can help overcome prediction ceilings associated with the base models, and provide users with improved predictive performance. These types of complex predictions would also greatly benefit from bias-correction techniques, especially Traditional Bias-Correction methods.

4.2. Techniques for Model Combination Model combination refers to blending or fusing the predictions generated by different base models, and is usually conducted at the decision or output levels - for instance, consider a scenario where several models predict the same meteorological variable, but at different spatio-temporal locations. The goal of the model-combination procedure would be to return a single value for the concerned spatio-temporal location at the concerned times, thereby condensing the

predictions from the individual models into a single prediction set - for example, on one model that may be at a different spatial or temporal location than the other models, the time zones could be different.

4.1. Introduction to Ensemble Learning

Ensemble learning is a meta-algorithm combining several supervised base learners, typically of limited complexity. The base learners could be thought of as groups of models capturing slightly different patterns from the accumulated experience. The basic idea is that a combination of weak learners with some degree of diversity will obtain a better generalization and decision boundary than a single and unique optimal classifier. The decision of ensemble learning is made according to a weighted (or unweighted) combination of the results of predictions given by the different models of the ensemble, which could depend on either the majority vote, averaging, exploiting rank, stacking, or others. The learning process can be conducted in a sequential manner such as in boosting methods or in a parallel method such as bagging or stacking. Ensemble methods are a special case of the broad topic of model aggregation or model combination [6,8]. The need for ensemble approaches arises from the fact that no single-learning algorithm has been found to be uniformly best across all applications. This leads practitioners to consider the use of sets or combinations of models in order to increase overall prediction performance. In typical usage, ensemble methods combine the predictions from each of the available models using methods that diminish the chance of an erroneous prediction by any one model in the combination. Techniques based on model averaging offer an intuitive approach to using collections of models. When individual models provide probabilistic responses, ensemble methods combine these probabilities in order to obtain an overall probability, based on which a prediction is made.

4.2. Techniques for Model Combination

Statistical modeling techniques to combine model outputs have long been used in the meteorological community. Among those are the linear or weighted mean, iterative methods like the modified Brier or logit correction, methods that use the rank order of the predictions, prediction interval calibration techniques, or combinations that are based on latent variable models. Other techniques, such as the best member from a multimodel, have recently gained prominence due to advancements in artificial intelligence and machine learning-based climate prediction frameworks. These regress and bias correction methods can not only add more value using the ensembles during the calibration phase, but also can address cases of catastrophic failures that can be found in the individual model results for specific variables, months, regions, or even years.

More recently, artificial neural networks have been proposed. Deep learning aims to model complex relationships from the data, and it is very encouraging that it has been applied successfully at multiple spatial and temporal scales to the post-processing of outputs from numerical weather prediction models. Even if theoretically an artificial neural network can model any combination of non-linearity and relation between input and output variables. The so-called "universal approximation property" of neural networks has not yet been tested on the extremely complex interaction of physical and parameterization processes in numerical weather and climate prediction models. Combining outputs from different models offers hydrometeorological decision makers the best of both worlds if it is done correctly, while avoiding the shortcomings of each individual forecasting/modeling system.

4.3. Applications in Climate Predictions

The most commonly studied problem in climate prediction is simulation of seasonal climate. The probability of satisfaction of a threshold value for a seasonal cumulative precipitation over a specified support has huge societal impact and hence it has been a subject of extensive research. The Multi-Model Ensemble approach has been used to confirm the increased predictability in spring predictability barrier using an ensemble of atmospheric GCMs. Similarly, the Multi-Model Ensemble methods also confirm the predictability of candidate seasons by using GCM monthly data for the Indian summer monsoon rainfall, Excursions of ENSO, MC, APC, IEO, and Atlantic Multi-decadal Oscillation. Multi Model Ensemble methods proved useful in prediction of dry/wet summer monsoon scenarios. Using forecasts and GMTs has shown skill in prediction of threshold crossing of the SWM seasons using methods for a 6 month lead time, and QM methods for a lead time 3-6 months.

K-category predictions have shown skill using KCM at one month lead time and at three months lead time using KENVA. Various models have been employed to predict a variety of climate data from El Nino Modoki and from sea surface anomalies. A K-category model has been employed at three month lead time for seasonal extreme predictions of the Northwest Australian region. The authors employed a K-category model at seasonal scale for categorical predictions for regions of North East India, with a one month lead time. The skill characteristics of MM-ENVA is suitable for Global Tropical Weather scenarios.

5. Uncertainty Quantification in Meteorological Models

Inherent uncertainty in physical systems and learning-based approaches motivates the quantification of prediction uncertainty in AI approaches. Modern meteorological models are built using sophisticated techniques, and are continuously improved and refined to provide pedagogical insights and enabling tools to meteorological scientists.

Experimental meteorological models based on AI models for scientific discovery can also support the scientific community with the unique challenge of being data-poor and model-rich owing to their basis in first-principles physics. However, neither set of AI models are reliable methods for making inference and predictions. Uncertainty quantification addresses this knowledge gap by providing predictions together with their associated uncertainties, which are useful to expert meteorology decision-makers, as they are trained to use and interpret both predictions and uncertainties. To ensure that both predictions and uncertainties are consistent and reliable, experimental and AI-assisted prediction/analysis weather models should undergo UQ.

Broadly speaking, UQ refers to methods for quantifying the uncertainties in the input and model parameters, propagating them through the system or simulation, and quantifying their effects in the output of the simulator. There are different flavors of UQ throughout the disciplines, which differ primarily in their application to either the input signal or the inference model. In meteorology, especially for the tasks of forecast or perform data assimilation, UQ for the whole model pipeline, from input observation to prediction or reanalysis, is of primary importance due to the inherent opaqueness of the traditional models operating in the background.

5.1. Importance of Uncertainty Quantification

Uncertainty quantification enables the generation of probabilistic models and is currently a major aspect of research in both the computational and scientific communities. Effective assessment, representation, and propagation of uncertainties is a crucial component for many significant applications, be it in engineering design and reliability, economics, molecular dynamics, risk analysis and management, traffic management, biomedical applications, or others. The desire to create surrogates based on physical, numerical/simulation models to provide fast, computationally efficient replacement models for complex, expensive computer simulations, has spurred the development of polynomial chaos-based methods. Although these approaches have been around in various forms for some time, the recent flurry of activity seems to be triggered by advances in high-order numerical methods combined with the widespread availability of high-performance parallel computing.

Uncertainty quantification also plays an important role in the field of meteorology and climate dynamics, although relatively few effects have been reported on the use of uncertainty quantification in operational meteorological modelling, as well as forecasts in terms of predicting large and small-scale features and for advanced warning systems. Weather prediction, and the ensuing warnings for potentially dangerous weather conditions, are based on the deterministic prediction of atmospheric states from numerical simulation models, which, to put it bluntly, render the present state of

the atmosphere and predict its future “weathering” on the basis of the developed thermodynamical principles that govern the evolution of atmospheric dynamics. Due to uncertainties involved in model parameterization, subscale physical processes, model configuration, and the chaotic nature of the atmosphere, such forecasts cannot provide an accurate representation of the atmospheric state or the true realization of the weather for times greater than a few days ahead.

5.2. Methods for Assessing Uncertainty

Quantifying uncertainty can be divided into two types: verification and validation techniques. The term "verification" is usually used to answer the question "did we build the system right?", while "validation" refers to "did we build the right system?" Verification measures how well the model reproduces fields that were used in its construction, while validation assesses the model results against other observations. Validation is usually the most critical measure used by experts to ascertain if the model is appropriate for prediction. In particular, for those who will make important weather-, atmosphere-, and climate- related decisions based on forecasts, the importance of accurate validation could not be overstated, because there are very direct and costly consequences for model failure.

While verification is more answerable, validation allows much broader conclusions called "well-suited-ness". Also, for complex, multi-parameterized systems, verification will have little generalizing power, because it must be done for all observed configurations. In practice, the verification tests do allow for wide-ranging conclusions, but they require the use of smaller reduced models. By using well-accepted reduced versions of the complex high-dimensional systems, it is possible to analyze the physical discretization instead of being forced into local behavior analysis as in full system verification. Nevertheless, any answer requires both verification and validation use, because both approaches provide complementary conclusions. In either case, validation will involve a more trusted multi-observation dataset and a designated period of time, typically using variable ranges outside those used for model selection.

5.3. Case Studies in Meteorology

In this last part, a few applications in meteorology are outlined that directly utilize inferences from statistical techniques, showing that probabilistic results are desirable and useful. Empirical results about such meteorological studies can inspire meteorologists and modelers to take a similar route because there are indeed benefits with the probabilistic perspectives. One reason that the results seem sound enough is that meteorology has plenty of data of excellent quality due to the support of microwave and infrared satellite techniques and the mass observations on the physical

states that spread in the recent globalization. The following list consists of dry-convective wind over Thailand and weather fronts observed by a satellite.

In the winter, a dry wind blows over the northeastern part of Thailand called as “Khamsamut" or “Krung." Using the synoptic and climatological features, an appropriate index is constructed to consider such dry days. However, the modeling part from the data is not so neat. There are complicated features such as zeroinclusions, marked seasonality, and heteroscedasticity. The modeling with appropriate UQ in the index enables one to understand each of those sub-regions better as well as to solve the inconsistency, which is important from climatological and ecological points of view.

The weather fronts that play important roles in temperature and precipitation, among others, are one of the meteorological features that have space problems. They need to be crisp to work for forecasters. In the last four decades, global data assimilation analyses have been produced as the product from the front model. By restricting to the limited areas using contouring techniques, the fronts can be used for deterministic forecasts.

6. Comparative Analysis of AI Models

In this section, we analyze and compare the models described and study their performance on different tasks by reporting various performance metrics. We list various strengths and limitations of these models in the table below. As seen from the results, many models outperform the others in specific tasks. Various models are best for data-driven tasks and some work best for science data-driven tasks, even though Transformer models manage to combine the advantages of both. Though models like CLIMB and U-Deep farming the new self-distillation and unified training approaches show significant improvements in these domains, it's always the more classic approaches like Precipitating Uncertainty, nowcasting and climbing uncertainty continue successively providing checkpoints on those tasks. It's interesting to see how different approaches tend to have different efficacies in providing results for diverse approaches.

6.1. Performance Metrics We compare various AI-enabled methods as follows, where we categorize them based on the problems they tend to solve. The methods differ in input modality and their final prediction targets. For example, one model predicts precipitation accumulation for a specific lead time, while another predicts blurry observations of precipitations. Then we specify performance metrics that have been reported within those specific problems. Given the reported metrics of different approaches, we utilize the standard metric per each problem as the base for our comparisons across different approaches. The majority of these metrics focus solely on evaluating the precipitation fields rather than real-world impact and applicability of the

models. Evaluating the performance score of each model inline potentially allows us to even better procedures. With such an evaluation, influenced internal structures or optimizable parameters for these models can be advised, allowing such minimal point-free subjective calculation – evaluating meteorological-driven or impact-based model designs easier.

6.1. Performance Metrics

Different ML models can yield varying performance levels for the same application. To appropriately select an AI model, it is important to use a meaningful metric; different aspects of a model's validity can be explored using metrics that reflect inflation or deflation of one or other inference dimension. The first step to measure the performance of forecasting models is the evaluation protocol selection. Its choice is crucial in order to assure that the evaluation will reflect model uncertainty properly in a forecasting application. QoF score is the simplest and by far the most widely used performance metric; at global level it is especially sensible to the bias error component. However, it is important to underline that the global definitions are overly sensible to the large scales events. It is understandable that they put more weight into larger scales, since big, important scales reside at the core of the meteorological models designs. Moreover, it is also important to stress that large scales are reduced in the low-miss Qof definition. IoU performs quite well at the object level of the different dimensions of interest, for objects of various size classes. QBS reflects that Qof is not sharply affected by underestimation errors. These last two metrics detect and compute precision and recall of useable objects in forecasts and observations, within tolerance, positioned around the observed position, at given scales. However, numerous critical features of the comparison are hidden in the above definitions.

The detection of the quality of the object predicted position, its pattern, its structure or its reasons still need further investigations. No single metric is able to catch all these errors. A more precise study is therefore needed in order to design a synthetic metric to even more directly express the model, or any other algorithm, inference quality. In other words, the degree of agreement between predictions and observations is far too intricate and important to be measured and expressed by a single number. It must also be stressed that, as far as synthetic metrics are concerned, the number of pixels affected by possible localized large-scale errors but also the specific nature of the considered forecast, i.e. decisive or deterministic, affect the evaluation. A more detailed examination can contribute to exploring different performance preferences, thus validating or invalidating more constructively models. The study of required adaptive/constructive biases as a function of model error is also important in order to improve inference performance and considering the AI model as a tool for forecasters.

6.2. Strengths and Limitations

The continual development of AI models for the tasks of meteorological data and climate pattern interpretation, analysis, modeling, and prediction has indeed been pursued on a wide scale with varied approaches to decision making for the design of the specific algorithms used. Despite this wide span of innovation across the disciplines and technical domains, one recurring and critical element of all methods is the careful calculation of their strengths and weaknesses for varied applications at hand. Models are trained and used for not only the domain specific tasks here but also on varied geochemical, climatological, astrophysical, and cyber-physical domain specific tasks with differing temporal-spatial correlations and resolution. While the general disciplines of physics, statistics, and even mathematics guide and mold this development, it is these domain critical features that decide the actual mechanism of model design as well as usage. It is these specific strengths and weaknesses of models that influence the decision scientists need to take when using it as well as policymakers in interpreting and accepting results.

In this chapter, we present a comparative framework for multiple AI models, their applications, strengths, and limitations towards clear and informed model selection for target datasets or specific tasks. Multiple AI models across classical statistics, data-driven models as well as physics-based methods are thereby analyzed and compared across common metrics but focusing on task and dataset specific features. We also discuss the design-specific, model-specific, as well as data-driven aspects of weaknesses and model constraints that can yield unexpected biases while using these models for accelerated discovery or revealing unknown geophysical information hidden within potentially multi-modal and multi-resolution datasets. Hence, informing best practices for scientific model usage and minimizing biases.

7. Future Directions in AI and Meteorology

7.1. Emerging Technologies Artificial Intelligence has driven and will continue to drive innovation in many fields, including, but not limited to, climate informatics, weather science, and weather prediction. Some of those emerging technologies include: Causal Inference: Including a causal inference framework in existing architectures, such as Score-based Generative Models or Diffusion models. This has wide ranging applications, including expecting extreme events in the future; understanding the causal drivers of singular events; reconstructing ancient climate conditions; and modeling of probabilities of occurrence for certain conditional distributions. Foundation Models: These models may infer important solutions like interpolation, climate reconstruction, and window-based weather prediction under large and rich prior distributions. Real-Time Deployable Models: Weather models are

run under limited spatial-temporal resolutions, as fast inference or prediction is critical. Here models that are also deployed in real-world energy applications such as solar or wind energy forecasts for optimization and decision-support are needed. Telescope Models: Where there are vast spatio-temporal-spectra-based high-dimension observations, telescope model platforms at either real-time or corresponding resolutions that model every sub-domain, by possibly sub-domain and fractal-infusion hierarchies, are sought. 7.2. Potential Research Areas Other research areas include, but are not limited to, the integration of multiple scales in AI modeling, extending existing state-of-the-art forecasting models in speed and training, long-term projection of short-term events, multi-modal learning, uncertainty quantification under High-Dimensional and Non-Gaussian data distributions and Emulation of Physics-Based Models with Artificial Intelligence.

7.1. Emerging Technologies

Recent advances in artificial intelligence (AI) and machine learning (ML) are revolutionizing science and engineering. Deep learning (DL), convolutional neural networks (CNN), and transformer networks are achieving or exceeding human performance for several challenging problems, including computer vision, speech recognition, and natural language processing. Although the science of weather and climate prediction is much older, the foundations of modern statistical meteorology were paved only three or four decades ago with focus on synoptic and mesoscale forecasting. Advanced models are now extensively relied on for prediction. The sophistication of the models is limited by the constraints of computational operation which has recently reached exascale levels. The growing availability of high fidelity data sets in terms of both global reach and resolutions, are fueling the development of more sophisticated ML algorithms. AI meteorology has recently emerged as an exciting interdisciplinary research area focused on exploiting those synergies. Some prominent research avenues have included the application of CNNs to nowcasting, hurricane tracking and intensity prediction, post-processing model output statistics with neural networks, and dynamical modeling with graph neural networks.

The recent advances, however, only scratch the surface. New technologies such as transformers, generative adversarial networks, attention-based models, large language models (LLM), and diffusion models, coupled with dramatic increases in computational capabilities and the growing wealth of high-quality big data, pave the path for setting a new paradigm in weather and climate applications. The emergence of large transformer models is transforming natural language processing, as evidenced by the performance of large models. Their astounding capabilities, including few-shot capabilities, zero-shot generalized transfer abilities, and human-like creative abilities

are sparking interest in the meteorological community for a potential transformer architecture for next generation meteorological models.

7.2. Potential Research Areas

Research opportunities in the application of AI in climate sciences range across different subfields. Improving the performance of long-established numerical methods for simulating weather and climate extremes and assimilating satellite and reanalysis data for these methods are open areas for collaboration. Coupled atmosphere-ocean-dynamics-vegetation-chemistry models that focus on a few of the dominant processes in order to avoid the high computational cost of full-fledged models for climate and weather simulation have made significant strides in recent years. However, there is still a knowledge gap to fill with respect to how to parameterize some of the neglected processes. For instance, a large fraction of weather and climate extremes is due to gas concentration in the atmosphere being above a certain tipping point, exceeding which the probability of these weather and climate extremes happening grows rapidly with increasing concentration. How to formulate these tipping points, as functions of other system parameters, in more intermediate and finer-scale models are questions that can be answered using machine learning.

Applying data-driven AI models to larger atmospheric flow problem by drawing again on reduced models is a research frontier. Another direction relates to how to use AI to improve the performance of existing weather forecasting models. Neural networks can be trained in a supervised manner provide guidance to those working on weather forecasting problems on how to choose parameters such as the search depth in multi-scale models. Physics-informed neural networks, which embed equations that describe the anticipated physical models in their feedback action, provide a mechanism for combining data-driven and physically grounded models. AI, therefore, does not lead to a displacement of government support for computational cost of forecasting the weather. Instead, AI could play a signified role in increasing the accuracy and scope of weather forecasts.

8. Conclusion

Climate Models can be improved in precious ways and our AI Models show their application in that. Quantified Earth System Models or emulators. AI tools can analyze fast the history of climate change on specific locations on the Earth and will provide spatial temporal visualizations of the results. Predictive Data can also include any non-physical related data, in theory, we would be able to build that type of prediction.

The climate is increasingly being used for private or government Institutions as a risk against natural disaster related projects, to spend fallen on one side of the equation.

They will give the value that haven't be considered for projects on climatological purposes like agriculture, health, and especially urban projects. In other words, our Models AI solutions for the Climate-Changes and Disaster are important to be found, for imagine safe projects.

Micro predictions but bring political solutions like Heat Wave Alerts and systems need to be generalized and followed by big initiatives in the area. Those information systems already exist; they will need to be used and easy to go along. AI technical tools like Machine Learning can be used for different kinds of predictions. By now we are developing Climentary Warehouse Capabilities. It's used in an online environment that is on Kivy and for big predictions online; we hope to per period by limiting dependences for Microenvironment Delta models. Focusing only on weekend predictions is also being studied.

In that spirit, we believe that in less than ten years, whoever can develop for urban regions a Safe-Life-After-Human-Main-Environmental-Actions Method, will be able to control what is to be Climate for Human future. Founders IA Statistic Tools say that we could in all the environmental life expectancy on every project, inside out.

References

[1] Abebe WT, Endalie D. Artificial intelligence models for prediction of monthly rainfall without climatic data for meteorological stations in Ethiopia. Journal of Big Data. 2023 Dec;10(1):1-5.

[2] Mu M, Qin B, Dai G. Predictability study of weather and climate events related to artificial intelligence models. Advances in Atmospheric Sciences. 2025 Jan;42(1):1-8.

[3] Talaat FM, Kabeel AE, Shaban WM. The role of utilizing artificial intelligence and renewable energy in reaching sustainable development goals. Renewable Energy. 2024 Nov 1;235:121311.

[4] Conti S. Artificial intelligence for weather forecasting. Nature Reviews Electrical Engineering. 2024 Jan;1(1):8-.

[5] Zhao T, Wang S, Ouyang C, Chen M, Liu C, Zhang J, Yu L, Wang F, Xie Y, Li J, Wang F. Artificial intelligence for geoscience: Progress, challenges, and perspectives. The Innovation. 2024 Sep 9;5(5).

[6] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.

[7] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.

[8] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.

[9] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[10] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[11] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[12] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[13] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[14] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[15] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[16] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.
[17] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.
[18] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.
[19] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.
[20] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.
[21] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.
[22] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.
[23] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.

[24] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.

[25] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.

[26] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.

[27] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.

[28] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.

[29] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.

[30] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.

[31] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.

[32] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.

[33] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.

[34] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.

[35] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.

[36] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.

[37] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.

[38] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.
[39] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.
[40] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.
[41] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.
[42] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.
[43] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.
[44] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.
[45] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 8: Satellite Imaging and Environmental Inference

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Satellite Imaging

Satellite imaging is an alternative source of information for evaluating various environmental variables and phenomena. Some advantages of satellite data over ground-based observations of similar variables include broader spatial extent, repeated temporal sampling, and more stringent measurement conditions. However, these advantages may come at the cost of uncertainties, due to lower spatial resolution, more complex relationship between the measured signal and the geophysical variable of interest, lack of measurements for some time periods, and calibration drift over time. Satellite data products have significantly advanced the fields of environmental monitoring and, importantly, evaluating the performance of environmental models [1-2]. A quantitative comparison with similar in situ measurements indicates the uncertainties associated with satellite data, including possible sampling bias and drift. Satellite instrumentation design has matured over the past few decades. The increasing number of satellites and sensors, greater measurement fidelity, and abundance of open data access have facilitated the application of satellite data in various disciplines across the world.

The first satellites dedicated to Earth imaging were developed in the early 1960s, culminating in a program that did reconnaissance mapping using high-resolution panchromatic photography [2-4]. In the following decade, several Earth observation satellites were launched that used multispectral data for land use characterization. The launch of both the LANDSAT-1 satellite and the Hubble Space Telescope represented a significant advance in Earth and astronomical imaging, respectively. The

hyperspectral remote sensing of terrestrial and oceanic surface processes was initiated with the launch of an airborne sensor in the early 1990s. Numerous dedicated hyperspectral sensors began operation before recent technology advances made hyperspectral instrument miniaturization possible for small and micro satellites.

2. Remote Sensing Analytics Using AI

2.1. Overview of Remote Sensing Satellite images have become ubiquitous in recent decades. There are a number of sensors, ranging from those designed for extreme resolution monitoring of tiny location of interest to those designed to be the eyes of the planet—with venerable histories of documentation of the Earth on 50 and 10 year time scales, respectively [5-6]. Indeed, several nations have also designed their satellites for similar purposes. Importantly, new sensors which use microwave, multispectral, hyperspectral, thermal or combinations of those wavelengths have also been launched to monitor specific physical processes of interest. For example, thermal images could be used to map water stress over huge distances when used over agricultural land. These satellites all have three characteristics in common: (1) they are optimized to monitor the surface of the globe; (2) they take images at periodically different angles to reduce problems of geometric distortion, particularly for large areas; and (3) they capture data at repeat time-intervals—generally daily, weekly, monthly, or once every 15-30 years—that sometimes lead to publicly available image archives.

2.2. AI Techniques in Remote Sensing AI techniques are ideal for the massive and complex data coming from the various satellite sensors. The data can be applied to interdisciplinary domains that address fundamental questions for policy-makers and scientists interested in monitoring land use or land cover change or any physical process on Earth [7,8]. For example, they can address questions such as how to cost-effectively monitor deforestation; how to optimally track the flooded area; how to best reduce algal blooms; how to design global coastal management standards for protection to identify safe harbors; and how to accomplish all of the data coherence—and more importantly science rigor—for a major international collaboration consisting of hundreds of linked science projects using Earth observing satellites. Policy-makers find these questions relevant because millions of people are affected daily by outcomes associated with hazards; provision of environmental resources; and provision of goods and services, such as risk reduction; and human, animal, or plant health. Indeed, there is now an explosion of interest in using AI approaches for applications in remote sensing that intersect environmental policy-making.

2.1. Overview of Remote Sensing

The process of quantifying the properties of a target–the measurement of the target using the interaction of energy and matter with the target–is called remote sensing. The

term ‘remote sensing’ itself originates from making measurements of the Earth’s atmosphere and surface without making physical contact with these targets [9-12]. Here, ‘remote sensors’ generally mean instruments employing electromagnetic waves (or signals) that are launched away from the sensor, hitting the targets, reflecting back to the sensor after interacting with the targets. Examples of invisible signals triggered by such sensors include active acoustic signals, and active EM as well as passive EM signals at infrared, visible, microwave, and radio bands.

Although the notion of remote sensing had begun in the 1930s, it was not until the 1960s, when remote sensing employing satellites, particularly those with optical, thermal, and microwave sensors operating at ultraviolet, visible, infrared, and microwave spectra, respectively, were launched, that the term was popularly adopted and widely accepted. Remote satellite sensors have by now been launched in their thousands, including thousands of ‘Earth observation’ satellites, with a wide variety of sensors employing several different spectral bands. With scientific objectives ranging from physical studies of planetary bodies to meteorology, geology, oceanography, limnology, ecology, and Earth surface physics/chemistry and their interaction with the climate change, many of these satellites with their onboard sensors are made available free of charge, and their data have been stored in large archives readily available to researchers worldwide.

2.2. AI Techniques in Remote Sensing

The AI revolution reshaping many areas of science and technology like natural language processing, computer vision, reinforcement learning, and multimodal and symbolic AI is now extending into remote sensing, augmenting many traditional approaches like hardware development, sensor design, image processing, computer vision, machine learning, and explicit reasoning about the physics of satellite sensors [7,13-15]. Nonetheless, within remote sensing applications the AI work is more focused around the domains of segmentation and classification of hyperspectral, multispectral, and RGB images, data fusion of multisource sensor data, product generation of surface reflectance, atmospheric correction, sensor calibration, and scene understanding via 3D reconstruction. The classical area of satellite and airborne sensor development, calibration, and image formation where AI is used is detecting deficiencies in the instruments that require interaction with human operators. Such AIs allow for a single line description of how to take a series of images, for example for sensor calibration that was normally done with a rich description using human coded software systems. Like with radio telescopes, the actual ground state product of satellite and airborne cameras are the collections of images and video, incorporating at best the initial instrument response function of the sensor, algorithmic solutions for sparse remote sensing of the geology and the land surface, and the actual information

latent in the collection of hyperspectral or RGB images which is comparing the collection against what is known or traditional AI techniques of least squares matching or optimization to minimize overall differences. Such AI systems are contributing to speeding up the process of mapping the globe and its changes from collection of images to site identification and prediction of land processes [9,16-18]. At a higher logical level, AI structures the observations of the Earth through the creating of learned databases of the Earth, only that the mapping of the globe changes as frequently as the observations using stereotypical methods for proving the existence of a particular kind of sensor measured object or process.

2.3. Applications of AI in Environmental Monitoring

Research in using AI in remote sensing has achieved remarkable success over the past decade. AI-based research in remote sensing can be broadly categorized into two domains [2,19-20]. Firstly, AI has greatly improved the accuracy of the majority of the common remote sensing tasks in the field, especially in the use of satellite images. Vision Transformers, Convolutional Neural Networks, Generative Adversarial Networks, Semi-supervised Learning, and Self-supervised Learning have all been adapted for improved performance in the remote sensing field, considering the specifics of remote sensing data. Secondly, AI has enabled new research questions to be addressed within remote sensing with the support of domain knowledge, especially new research questions that go beyond the standard paradigm of single-image single-task learning. This includes multi-dimensional and multi-task learning with extents along both the spatial and temporal dimensions.

The use of AI in remote sensing has been the enabler of many successful satellite-based environmental solutions. The temporal archives of the observations that can be collated in satellite space weather stations provide additional data resources that allow the use of the AI at two major levels of insight [9,21-23]. At a lower level, such as in vegetation disturbance detection using pixel-wise inversion of change detection, AI post-processing of lower-level remote sensing products such as change maps has been the norm. At its higher level, AI is being deployed to directly monitor key elements in Earth's energy exchange equilibrium climate system at a sub-continental scale. The major monitoring areas include land cover classification and monitoring, land cover change and vegetation disturbance detection, vegetation biophysical assessment, and radiative transfer modeling for flux estimation. Other monitoring areas also cover water body monitoring, geopolitics analysis, air quality monitoring, solar flux and radiant energy estimation, and smoke tracing for surface water quality, toxic releases, and impact on human health estimation.

3. Land Cover Classification

3.1. Importance of Land Cover Classification

Land cover classification from Earth observing satellite sensors, in either full hyperspectral datasets or reduced multispectral or broadband data, has become one of the standard products available from satellite missions. Most notably, products for classifying pixels into land cover categories such as vegetated areas, barren rock or sand, water, and urban areas have been produced for nearly all of the Landsat missions and over large areas for the National Land Cover Database. Increased interest in monitoring change in carbon fluxes globally has highlighted the need for products that go well beyond standard land cover classification schemes to classify landscapes into their structural and functional components, e.g., carbon-dense intact forests, degraded forests, croplands, and shrublands. In this context, investors in satellite technology are producing new sensors with unique capabilities.

3.2. Methods for Land Cover Classification

There are two general approaches to land cover classification. Empirically based statistical classifiers, such as Gaussian maximum likelihood, k-nearest neighbor, and Parallelepiped classifiers, among others, learn template spectra from training subsets of the data and apply these classifiers over the entire scene. Decision tree and random forest classifiers based on increasing classification accuracy by sequentially partitioning feature space or combining the votes from classifier trees also utilize training sets. The second class of classifiers utilizes methods such as neural networks, support vector machines, and deep learning techniques, which incrementally increase the accuracy of the classification from iterative training of many prototypes, each based on varying weights that sequentially mimic human decision-making criteria without requiring explicit templates from users.

3.3. Challenges in Classification Accuracy

Regardless of the method used in the training phase, the ultimate question is whether the classifier accurately reproduces known categories for a different validation set of data at locations where land cover is clearly defined. Often these validation locations are limited. Results can be strongly affected by classification methodology, timing, illumination, view geometry, atmospheric conditions, land cover seasonality, cloud and shadow masking accuracy, and the location of training sites used for supervised classifiers.

3.1. Importance of Land Cover Classification

Land cover classification aims to automatically partition Earth's surface into discrete land cover types such that pixels or objects in the same category behave similarly according to some criterion. Land cover classification is the backbone of natural and artificial resource management, disaster impact and recovery assessment, climate change and health-related studies, and any practical application where ground truth is difficult and expensive to collect, but in general needs qualitative assessments of the same or similar categories over large areas [24-26]. Such descriptive information includes vegetation type, volcanic cover type, soil erosion type, ground water quality, terrain slope type, road or pathway type, as well as flooded or burnt areas, and their relationship to climate, geology, demography, economy, health, and other factors. These categorical values are themselves observations, but the combination of the original matrix with the value maps forms a physical model of the type, providing further information that is not usually available from the original sensors alone.

Land cover type mapping and monitoring are also useful for helping to understand the temporal and spatial evolutions of related biogeophysical processes. For instance, the seasonal changes in land cover condition may help models explain the seasonality of terrestrial carbon fluxes, and the increasing duration of snowcover or frozen condition over forests or tundra may help explain the observed increasing boreal complex carbon fluxes, and the increase in the annual number of hot days and the consecutive dry days in a summer season over a shade tree growing urban may help explain the increasing peak urban carbon fluxes. Addressing these problems requires long-term carbon flux observation networks, which are currently not possible due to the lack of land cover type mapping using the same or similar techniques for a long period.

3.2. Methods for Land Cover Classification

Land cover classification methods are broadly categorized into supervised and unsupervised approaches. Unsupervised classification seeks to classify pixels into groups that are spectrally similar; it is non parametric with no previous knowledge of potential land cover classes. Common examples of unsupervised techniques used for land cover classification include the K-means classification where K cluster centers are initialized with random pixel values belonging to different classes. The classification proceeds by assigning each pixel to its closest cluster center and computes new cluster centers, repeating this process until the sum of squared distances of all pixels to their corresponding cluster centers converges [8,27-30]. The IsoData classification expands on this by allowing the user to specify thresholds for convergence and a maximum number of clusters. The unsupervised approach may work well in situations where the

image contains a large range of unique pixel spectra; however, it often fails to attain the desired classification accuracy.

Supervised classification requires the user to generate training sets containing spectra from example pixels representing each land cover class prior to executing the classification process. Samples are often verified post classification for accuracy. The user also requires previous knowledge both about which classes may be present and what spectral characteristics define them. Statistical supervised classifiers use statistical methods as a guide to classify each pixel based on its spectral signature. These classifiers include the maximum likelihood classifier, the Gaussian maximum likelihood classifier, and the minimum distance classifier. More recent classifications, however, leverage a variety of machine learning algorithms such as random forests and convolutional neural networks for high-dimensional data. Often, these machine learning classifiers can achieve much higher accuracy than their statistical counterparts, particularly for high-resolution, high-dimensional data. However, developing an accurate machine learning classification involves a systematic search through algorithm parameters, including the number of trees as well as loss and activation functions, which may not always yield optimal results.

3.3. Challenges in Classification Accuracy

Classification is important for many applications, such as mapping urban extent and change, mapping vegetation and soil types, and mapping the extent of agricultural areas and crops. Accuracy is central to this process, whatever the classification method. Here we discuss some of the issues that can lead to uncertainty in classification [9,31-33]. The first and perhaps most obvious reason is the limited spatial resolution of coarse resolution satellite sensors. Classification accuracy is compromised when users are interested in small targets. These requirements suggest the need for high resolution or very high resolution satellite images.

A more subtle reason, however, is that themes classified by users are employed as “truth”. Thematic accuracy depends to a much greater extent on the difficulty of actually performing the classification and degree of consensus between the experts actually carrying out the classification than on the analyst's choice of classifier, which is typically the focus of accuracy research. For the maximum likelihood classifier for example, which assumes multivariate normality of categories in data space, it is generally agreed that the classifier works poorly under highly contaminated conditions. How much higher will classification accuracies be if the correct classifier is invoked than if not? No well-established theory informs users about this central issue. Answers may have to come from simulations.

4. Normalized Difference Vegetation Index (NDVI)

Several well-known products used extensively in various research fields apply the mathematical model of vegetation index. The NDVI, a numerical indicator that uses remotely sensed data to assess whether the target being observed contains live vegetation or not and it works by observing the different wavelengths of sunlight that are reflected by plant leaves. The NDVI produces a value between −1 and +1. NDVI is one of the most popular and frequently used remotely sensed vegetation indexes due to its high correlation with vegetation biophysical parameters and sensitivities to atmospheric and soil factors, ease of calculating NDVI with many satellite data in various land cover applications, wide application history of NDVI in many regions and climate regimes. Various researchers have reported significant correlations of NDVI with Leaf Aroma Index, Fraction of Absorbed Photosynthetically Active Radiation, Gross Primary Productivity, Plant Height, Vegetation Water Content.

4.1. Calculation of NDVI

NDVI is computed as the normalized difference between spectral reflectances for the near-infrared and visible bands: NDVI = (NIR−Red) / (NIR + Red).

It is found that NDVI ranges typically between −0.1 and +0.9, with most vegetation types falling between +0.1 and +0.9. Generally bare soils have NDVI values of about +0.1, whereas healthy dense vegetation has an NDVI value near +0.9.

The underlying theory behind the NDVI computation is based on the distinctive spectral signature of chlorophyll, which absorbs most of the visible light in the blue spectral region for photosynthetic vigor, while reflecting strongly in the NIR spectral band. Vegetation can be effectively discriminated from other nonvegetated surfaces because these two regions of the electromagnetic spectrum are spectrally different.

4.1. Calculation of NDVI

Of the many spectral indices published since the advent of satellite imaging for tracking Earth surface changes, the normalized difference vegetation index (NDVI) stands out as the preeminent, ubiquitous, and oldest. It remains at the forefront of all applications of spectral indices with no indications of decline in interest. In its most recognized and common formulation, NDVI is derived from reflected radiation in the red (R) and near-infrared (NIR) bands:

NDVI = (NIR - R) / (NIR + R).

NDVI varies between -1 and +1 though values closer to zero indicate no vegetation and higher positive values indicate greater cover and/or health of vegetation. NDVI

should be calculated from surfaces corrected for atmospheric interference as well as for the effects of viewing (or sun) geometry using calibration and/or correction models where available.

Because of the relationship between NDVI values and chlorophyll levels, NDVI is sensitive to vegetation removal, such as land conversion, development, agriculture, and fire, as well as stress and senescence. At the other end of the scale, NDVI has productivity and carbon cycling use in cases of afforestation, reforestation, and changes in the length of growing season, as well as global scale use in determining the relation between NDVI and interannual variation in carbon cycle-biomass activity. Data from NDVI calculate numerous environmental questions surrounding relationships among climate, temperature, carbon, and phenology.

4.2. Applications of NDVI in Environmental Studies

The NDVI has numerous ecological and environmental applications. With its strength in distinguishing between vegetation and soil–water content, the NDVI is a very popular and powerful tool for monitoring and studying vegetation dynamics and growth. These include mapping vegetation cover, leaf area index, chlorophyll content, carbon fluxes, drought and desertification monitoring, integrating biophysical and biochemical variables, and studies related to flora and fauna [34-36]. To this end, NDVI is extensively used to relate indices from optical remote sensing to variables that are otherwise measured using ground–based instruments. As NDVI is easy to compute from many satellite and airborne missions, and as remote sensing is mostly done at about the same time as field site observations, NDVI provides a suitable method to correlate the field variables with observations from satellite or airborne platforms. Once a valid relationship is established, NDVI is further used as an operational tool to map and monitor the changes in vegetation growth and dynamics over strikingly larger areas that were not feasible using the ground–based lens and instruments.

The NDVI can also be very effectively used to assess and map vegetative–cover stress, be it natural or manmade [3,37-39]. For instance, NDVI can track the historical patterns of agricultural drought and famine in different areas over time to find links between crop failure and population migration or death. NDVI is also used in climate and weather studies related to rainfall, temperature, water vapor, deforestation and its consequences, fluxes of carbon dioxide, and so on. NDVI is used in ecological modeling related to defining fire environment, fires and wildlife habitat, or global ecosystem changes. NDVI can also be used in hydrology to better model evaporation and ground temperature parameters by using them as inputs.

4.3. Limitations of NDVI

On NDVI's fringe, especially on the low end and in specific regions, multiple low-reflectance stimuli are possible. In sparse areas, e.g. in some drylands, such low NDVI may be caused by non-vegetation factors like the brightness of bare-soil, sand-dune surface, or dry grasses. The low NDVI values of some areas with dense Albizia and other broadleaf canopy trees include factors like the small leaf area index or the distinctive leaf spectral characteristics of these trees, or even that of certain tree barks. Some green-crop regions of dense but short vegetation may also show low NDVI reflectance [36,40-42]. In these circumstances, soil-adjusted NDVI may help NDVI to define vegetation cover. However, the need for soil-adjusted NDVI indicates a limitation in NDVI application.

On NDVI's other fringe, especially on the high end and in specific regions, a high NDVI is also ambiguous, meaning that not only dense but also certain irrigation or subtropical-plantation rainforest crops with light-green foliage could produce somewhat also-high NDVI values. Similarly, NDVI is not able to indicate crop types or tree canopy characteristics if it is used alone. NDVI-derivable greenness is ambiguous because NDVI is unable to discern the difference between reflected wavelengths that NDVI equates at a specific reflectance range. NDVI attributes such greenness difference not to reflected wavelengths' different spectral characteristics but to a relation, that is, a difference in proportion, between reflected wavelengths at an NDVI high or low. As a corollary, NDVI would not be able to distinguish between tree coronas or canopy foliage at various growth stages which have similar visible-light-derived greenness NDVI values but different infrared NDVI values. Consequently, NDVI might not be able to detect botanical stresses of plant crowns. As such limitations suggest, NDVI should be applied with care.

5. Emissions Tracking

Emissions of greenhouse gases (GHGs) are the result of the combustion of fossil fuels. The energy required to drive the engines of vehicles, to heat buildings, and to convert exothermic reaction precursors into cement, mainly derived from fossil sources, is taken from the burning of coal, oil, and gas. These emissions have led to the increase of atmospheric concentrations of GHGs. Monitoring the rate of emissions in the world is a rather technical but necessary task for validating pledges made as part of the Paris Agreement. The data reported by countries are updated infrequently, for example inventory data for 2020 were reported in mid-2022, and use proxies relying mainly on bottom-up approaches and sophisticated models. These models are fed with data from national statistics offices and adapted to known questions concerning emission outputs. The accuracy of those estimates is greatly debated in the scientific literature. As for

climate and weather monitoring, a merger of bottom-up approaches and top-down approaches using satellite imaging technology is more suitable for local and regional inference and makes it possible to strategize the path towards reducing emissions. The main goal is to reduce errors in both modeling approaches and establish a memory of the observed GHG concentration time series for the calibration of the models.

Estimates of GHG concentrations and emissions could be made using both passive and active optical sensors deployed on aircraft and satellites or ground platforms. In this chapter, we are specifically interested in satellite imaging techniques as part of the top-down approach [40,43-44]. Carbon dioxide and methane, the main GHGs, have well-defined electromagnetic absorption spectra in the optical and infrared ranges. Detecting their presence in an air column is made possible by satellite-borne sensors capable of measuring the top-of-atmosphere reflectance and radiance. A GHG concentration anomaly signal will be detected relative to direct surface reflectance for the reflected radiance and after correcting for surface temperature for the emitted thermal infrared radiance.

5.1. Techniques for Emissions Detection

Various techniques have been proposed to detect emissions using satellite images. Optical remote sensing—the most commonly used technique—measures the radiance that is reflected from or emitted by the target. Optical imaging is mainly used for land cover mapping or for measuring spectral reflectance. Because optical remote sensing is passive, it does not involve any interaction with the land covers. However, optical remote sensing has inherent drawbacks. For example, it is not effective in cloud cover, and it only works during the daytime. Moreover, it has the sensor noise problem associated with high-resolution optical satellite images, namely the sensor-generated noise on the high-frequency details of the optical images.

Radar remote sensing is another well-known technique using synthetic aperture radar. It is an active microwave imaging tool that has the time and wave frequency band flexibility and can perform imaging during the day or night or through the cloud. Such advantages of active sensors make radar attractive imaging tools in various applications in a variety of fields. The electromagnetic waves transmitted by the radar penetrate the atmosphere and interact with the land cover surfaces as they return to the sensor.

Multispectral thermal infrared images from remote sensing satellites are powerful tools for a variety of environmental applications. Many empirical observations as well as methodological advancements have been reported to obtain estimates of land surface temperature from the historical data of various satellite programs. However, satellite data-related land use and cover change studies have not yet used land surface

temperature data derived from the above sensors. Moreover, some satellite programs to develop land surface temperature data have previously been focused on detection of land cover changes over areas of the Earth's land surface.

5.2. Role of Satellite Imaging in Emissions Tracking

Accurate identification and tracking of emissions sources play a critical role in climate change mitigation initiatives. The monitoring of methane emissions sources is key to transitioning toward lower methane emissions [7,8]. As the global economy emerges from the COVID pandemic, enormous interest is being focused on how traditional carbon emitters can return to pre-2020 levels, if not increase their outputs. Regular satellite observations of the world's largest coal-fired power plants would provide a considerably enhanced capability for tracking carbon emissions. Recently, debates have ensued over whether or not governments should be given advance notice of an upcoming satellite mission when national security or surveillance data is concerned. These activities and events need to be followed for ongoing compliance with climate change terms for countries under specific agreements.

In the aftermath of Russia's invasion of Ukraine, a further surge of interest in satellite imaging for gas flaring emissions detection is being prompted to recognize security implications of satellite images of gas flare emissions associated with oil production. To this end, a technology company announced that it would focus on the continuous satellite monitoring of global gas flare emissions as an added security enhancement. Corporations throughout the world will be held responsible for compliance with the new sustainable development regulations that are being discussed and implemented. Therefore, in order to assess true growth in corporate value, as well as governmental adherence to climate change protocols, satellite monitoring of emissions will play a critical and ongoing role in the future with respect to environmental safety and security.

5.3. Case Studies on Emissions Monitoring

Emissions sources traditionally tracked by ground-based sensors have proved too distanced from urban pollutions for application of the local assignment models. This should no longer be the case with the capabilities of distributed sensors provided by large satellite constellations, making those sensors invaluable for testing remote methods against numbers published by the countries' monitoring networks. We illustrate such small satellite capabilities on the example of nitrogen dioxide (NO2) emissions tracking for Europe. With such experimental capability check, results for the NO2 fluxes derived from satellite sensing with the use of local assignment approach are compared to the results obtained by the use of the same local assignment method

but at local scale with the measurements from ground-based Continuous Emission Monitoring Devices.

According to this available validation, compared NO2 emission patterns for individual countries roughly correspond to the patterns based on CEM sources, properly noting differences at the influence of specific events, like volcano eruptions and COVID-19 pandemics on the scale of the entire continent. This and the data processing limitations due to the distinct spatial footprint of satellite sensing calls for the satellite determination of aggregated country-level NO2 trends rather than statements about the precision of satellite estimates. Therefore, publications from the scientific community are needed to fine tune proper scaling at least for the period when both ground and satellite data are available. This is also true for CO2 flux assessment, which still requires a few alternative solutions at least at the global scale prior to remote but still satellite-assisted applications to our domestic city.

6. Fusion of Optical, Radar, and Multispectral Data

The fusion of multispectral and optical data with older, but higher resolution radar and optical data allows scientists to take advantage of all three sensors strengths to maximize spatial, spectral, and temporal resolution, while minimizing cost. Optical and multispectral imagery, especially at higher resolutions, are ideal for monitoring spatial variations in environmental processes, but are limited by their temporal resolution because cloud cover can inhibit their use. Radar data, especially of lower spatial resolution, is much less limited by cloud cover and have it normally fall off at the perspective angles used by many time-series satellite image acquisitions. Fusion of optical data with other sensors to reduce limitations of individual sensors are common and might be used to fill in atmospherically induced gaps in optical data.

Sensors use image bands for cross-sensor correction, data merging, and calibration for multi-sensor applications, while others are often used to perform distortion errors for fine scale applications. The need for accurate atmospheric correction and geometric registration between sensors would require use of automated and advanced calibration methods. Radar and lidar systems have space-borne systems which are often fused with other optical weather satellites. Data are fusing data from both sensors to obtain enhanced normalized difference vegetation index integrated within a geographical information system for spatial and temporal dynamic modeling of wetlands and vegetation cover.

The higher spatial resolution of the optical/IR data can provide a better estimate of within-pixel heterogeneity and its longer wavelengths allow an accurate estimation of soil properties such as moisture conditions and mineralogical composition. Recent studies have suggested that data fusion to improve the availability and quality of

landcover classification and moisture monitoring has sufficient promise for better specification of heterogeneous pixel models and building.

6.1. Benefits of Data Fusion

There is a wealth of optical, radar, and multispectral data available for Earth imaging, with long archives for optical and for some radar systems. What we present here is, in effect, a how-to-make statement for visualization; it will be difficult to make a strong inference from a composite image of different modalities without some understanding of how the data may, and may not, respond, and how they interact. The broader vision, however, is of exploiting higher dimensional data from optical and multispectral imagery, their temporal evolution, depths from photogrammetric methods, and excursion into radar and lidar domains, for pixel- and object-based classification of earth's surface components and physical, chemical, and radiative transfer processes.

Data from different sensors and imaging modalities complement and enhance one another. We exploit the phenomenology of the optical, multispectral, radar, and lidar signatures of the various objects on the scene, recognizing their strengths and weaknesses; at the same time, we learn from empirical training and from physical models of signal propagation and transformation through the atmosphere, to optimize the imaging- and physics-modeling approaches. Impact of fusion on applications for these are manifold. One intrinsic challenge is to maintain the language and the congruity of dimensional size between coarsely pixelized and finely pixelized datacube components. Overall, data fusion addresses some of the most critical needs in modern remote sensing: classification and change detection from lighting conditions, techniques, diurnal and seasonal constraints on the availability and validity of specific data types; maximizing learning opportunities rather than waiting for optimum sensing conditions and resources; improving the accuracy, correction, and uncertainties of outputs at high levels of specificity, resolution, and reliability. Data fusion expands the domain of usability. Where numerous techniques apply to multispectral and optical data for issues of high specificity, general functional relationships, and canonical approaches need to be developed with regard to radar and lidar.

6.2. Techniques for Data Fusion

There are several approaches to data fusion, including those that are acoustic-linked or optical-linked, as well as empirical models, mixture modeling, texture, varietal models, and pure linear combinations. Acoustic-linked or optical-linked fusion assumes that penetrometers detect properties or layers in soil that affect spectral response to optical sensors, and vice versa, thereby further identifying suspected layers in fusion. In the models, one model detects the soil/mechanical properties that best explain the other

model's data, and the other model is fit to the penetrometer-derived soil/mechanical properties. Next, hybridization assumes that layered models best reflect the actual nature of the microstructure, but may or may not be detected using a limited subset of applicability of model optics.

Empirical models construct a model optical response to a known ground signature of soil, mechanical or otherwise. The mixture model assumes that the optical signature is created by two or more optically active materials or products in a pixel. This requires that multispectral imaging identify an area large enough to minimize spectral noise. Fractal geometry deals with the propensity of all surfaces on Earth to yield correlations of surface microstructure that are constant with scale. The techniques mainly rely on texture patterns, which include roughness, color, features, and microstructure enhancement, such as edge, corner, or feature detection.

The no-hypothesis model is a pure linear combination optically-based model that is valid for any combination of genuine optical observations whose pixel dimensions are larger than or equal to the largest structure of silhouette, in the receptive field of the imaging instrument in terms of spatial location and direction of view within the illumination cone, and radiance properties, using the discrete Fourier transform and a specific screen model. According to the model, a point source illuminates a stack of homogeneous planar screens that are opaque for infrared but transparent to shorter wavelengths. The model leads to a very informative and reliable tool for estimating instantaneous intense hypsochromic and dependent radiances of the solar rayleigh scattered light. The model has also been used to derive a simple formula for the true coherence.

6.3. Applications of Fused Data in Environmental Science

The complementary strengths of radar and optical sensors extend the possibilities for applications of fused data to new dimensions. Although the applications landscapes of optical and radar systems are both broad and deep, there remain many environments and phenomena that can be described in greater detail or with more reliable information content when both types of data are leveraged together in a joint information system. The advantages of fused optical and radar data in enhancing understanding of the physical processes that govern scenes on the earth surface are apparent in most areas of critical environmental interest where optical and radar observations have been used independently, including studies of vegetation cover, surface waters, urban environments, and snow-covered areas. The advantages are realized via improvements to key limiting factors on optical imaging, and those affecting radar remote sensing.

In terms of natural or built surface properties, radar remote sensing can provide some information about surface geometric characteristics and, for coherent systems with sufficiently high temporal radar resolution, surface motion and land surface displacement due to anthropogenic or tectonic subsurface stresses or geophysical stresses related to surface hydrology. Enhancement of retrievals and estimates made using either optical or radar data has been demonstrated through various data fusion techniques.

7. Case Studies in Satellite Imaging

This chapter explores a few selected studies using satellite imaging applied to environmental inference. The selection mirrors my research interests over my career, appreciation of the quality and relevance of the studies, and variety in the applied remote sensors available, applied algorithms, and inference problems targeted. Beyond technical advancement oriented research questions, an ongoing challenge with the applied use and evaluation of remote sensors is the limited number of in situ measurements available to develop and evaluate inference methods. Different from in situ networks that have been maintained for many decades, taking specific measurements in a certain location at a certain time, in situ “snapshots” of a specific temporally correlated event are difficult to obtain. Consequently, the main objective with the case studies here is to illustrate the representative potential of satellite imaging for environmental inference, and its technical implementation challenges. First, to monitor urban development in the city of Shenzhen, a multi-resolution study using a diverse sensor set exemplifies the unique sensor capabilities available to level the tradeoff between temporal and spatial resolution with different sensors, and the use of an off-the-shelf land-use classification algorithm to successfully relate the satellite image features with the specific event at hand. Second, an example of the use of unmanned aerial vehicle based imaging applied to deforestation tracking from an in situ perspective illustrates a challenge with using low resolution satellite imaging, and the potential of combining artificial intelligence augmented UAVs for rapid acquisition and evaluation of in situ snapshots with satellite imaging for extensive monitoring. Third, urgent environmental problems – in this case freshwater ecosystem health and socioeconomic effects of harmful algal bloom events arising from freshwater eutrophication – are highlighted as targeting a set of spectra relative to specific biologically or chemically active substances distinct in their fluctuation timing, but similar in temporal correlation with satellite imaging, is critical to data selection and quality for successful event response prediction.

7.1. Urban Development Monitoring

Cities are complex systems developed according to economic, social, political, and cultural characteristics of the area where they are located. Satellite imaging is a very useful tool for urban studies, with several advantages: the views from the satellite perspective offer an overview of the urban area and its surroundings, with the possibility of capturing a large coverage area and repeating acquisition of the same region over time. This large coverage area, coupled with the low cost of a large number of data collections, allows monitoring urban dynamics in a more robust way than results from small scale studies, in situ collection efforts, or even UAV data collections. But it is important to consider that, especially for urban analysis, the quality of the data and the spatial resolution needs to be appropriate for the intended applications. Current long term space missions are producing large quantity of satellite data, including some with high spatial resolution. A range of applications have emerged for the combination of these data with in situ observations to assist decision makers in city management and urban/spatial planning.

Urban development may have productive externalities and generate benefits to the surrounding locations, as a driver of population or market opportunities. However, excess demand for the labor market and other services required, when imbalances occur, might lead to depressed characteristics and growth impediments, such as urban poverty, slum, and informality. In this sense, it is important to devise indicators for monitoring long term urban development, on the identification of not only the pace of change, but also the direction of growth or decline over time. In this paper, we show an example of the type of indicators generated using satellite imaging data, as zoom proxies for the level of economic activity in the region.

7.2. Deforestation Tracking

Deforestation is a serious concern worldwide. Not only do large forest areas accommodate a variety of wildlife, but they also act as carbon sinks, absorbing more carbon than they release. Hence, there is a large number of environmental initiatives to prevent deforestation. Monitors keep track of forest loss. Satellite imaging provides an efficient means to remove contrasting reflection values of forests, which are spatially coherent, from noisy aerial images and helps global initiatives that try to counteract the negative consequences of deforestation. Light detection and ranging enables measurements of light at high points and provides elevated view special resolution ground based images, helping to develop the understanding of the limited type of reflection properties that the variety of objects in our world, such as human made structures, offer. The assumption of the existence of reflection values that vary little over a period of time enable detection of changes in reflectance of an object’s surface

which correspond to the object being removed or remaining. While monitoring of urban areas has received sufficient attention, not much focus has been given to observe deforestation and its ecological impact with these special images collected over time and apply inference metrics on those images, which enable the increased speed of inference analysis.

7.3. Water Quality Assessment

Water quality conditions are essential for ecosystem health and degradation and can have profound implications on economic production in coastal areas. For many coastal states, fisheries account for a large fraction of GDP, and large numbers of tourism-related local industries are also heavily dependent on healthy aquatic ecosystems. For these and many other reasons, monitoring studies of water quality metrics, such as chlorophyll-a concentration and turbidity, are important for sentinel monitoring tasks. For detecting broad-scale temporal and spatial trends in water quality, remote-sensing-based studies are especially useful. Historically, achieving the required accuracy when deriving water quality metrics from optical remote-sensing data has been challenging due to atmospheric interference at the low wavelengths where the ocean color response to the water quality metric is most sensitive, as well as the complex optical interactions in the top layers of the ocean at reflectance values typically seen in coastal and inland applications.

As satellite sensor resolution improves, more and more studies are using spatially disaggregated, midresolution optical satellite data to derive temporal descriptions of water quality parameters in coastal and inland waters. Here, the main focus will be on reviews of some recent applications of these midresolution, multispectral satellite data for chlorophyll-a and turbidity mapping in that these are the primary metrics of interest in operational monitoring by coastal states. Although these operational monitoring activities and many research applications use empirical and semiempirical, signal–metric inversion schemes to derive chlorophyll-a and turbidity concentrations, it must be acknowledged that these schemes are essentially multiscalar, using information at the different scales of the component bands and band ratios. There has been recent activity in the use of EOF-based multispectral approaches for coastal inversion problems, but there is still relatively limited activity in the research community. The primary challenge until now has been the missing data problem, which obviously is dealt with in the univariate waterfall and multitemporal plankton bloom studies.

8. Future Trends in Satellite Imaging

Satellite imaging presents new possibilities and applications, especially when combined with internet of things devices that monitor various elements and phenomena from the ground level. With the looming challenge of data abundance

beyond multi-terabyte satellite datasets and large-scale creating connecting from land-, sea-, and space-based IoT devices, coupled with decreasing satellite revisit cycles that capture millions of satellite images yearly, the quest for real-time ubiquitous intelligence drives the next wave of innovation on small, agile satellites. Small satellites and mega-constellations speed up the temporal revisit time of the Earth while new satellite technologies enable the capture increasing resolutions of satellite data such as hyperspectral and thermal satellites, radar and synthetic aperture radar imaging, and light detection and ranging imaging. These developments ready the stage for novel satellite devices, architectures, and constellations that enable new scientific investigations and inquiries.

Traditionally, remote sensing and satellite imaging algorithms have been in development since the 1980s and have focused on handcrafted models with prior knowledge of the features of interest. While these algorithms have been highly effective, recent advancements in AI and big data present new opportunities for continuous or continuous learning on the applied models that inform the sampling and fuse the in situ and satellite-level groundtruths in real-time or in temporal drifts such as seasonal-and annual-based integration of the physical knowledge and statistical distributions. The rapid breakthroughs in AI and machine learning capabilities have transformed our distant view of the Earth and revolutionized many areas of research. The pace of innovation has been rapid in many subfields of ML, particularly in private and commercial sectors, where data and innovative uses have converged to create new private sector capabilities and services for industry and government.

8.1. Advancements in AI and Machine Learning

With an ever increasing volume of satellite data being generated, the requirement for automated solutions to tackle conversion into actionable knowledge is essential. Coupled with an extraordinary growth in the application of machine learning techniques, particularly deep learning, the discussion around artificial intelligence has moved to the practicalities of its use. Exciting advances are being made in many problem domains such as target detection and classification, scene classification, segmentation and change detection, and in the use of satellite imaging in specific application domains such as forestry, environment, and urban development. The use of AI enables an enhancement of the capabilities of shortwave infrared microbolometers, which can be applied to identify agricultural products and environmental pollution more accurately and with better efficiency. Following on from these specific explorations, what has been lacking is a generalized overview of the role of AI in satellite imaging and inference, covering both information extraction from raw satellite imagery and guidance of the imaging process to maximize the efficiency and quality of imagery and inferred information. In this chapter, we report on such explorations,

highlighting many of the significant advancements that have been made, providing a digest, categorized by application domain, of those areas where AI has been used to good effect. Through this discussion, we identify places where current developments are lacking, and hence capable of being further exploited in providing solutions to untackled problems, and areas where better or more efficient AI solutions could be applied to improve existing results. Importantly, for many of the application domains, AI can enable solutions to be operationally viable, with the work providing an extensive resource for many taking those first steps into exploration of the use and applicability of AI in the context of satellite imaging and inference.

8.2. Emerging Satellite Technologies

Traditionally, Satellite imaging has been dominated by a handful of public and commercial agencies who build and operate large, sophisticated, and expensive satellites and provide infrequent, low-spatial resolution, low-noise data products. However, in the coming decades, novel satellite technologies are emerging that promise to change this orbit. The rapid miniaturization of sensor and communication technologies is leading to the deployment of large, distributed swarms of micro-nanosatellites equipped with visible, infrared, thermal, and radar sensors. These constellations will provide unprecedented global coverage at coarse, moderate, and high spatial resolution, enabling near real-time monitoring of critical events around the world. These systems will be coupled with advances in global positioning and communication technologies that will result in rapid on-demand data acquisition of any site on the globe. Capable of carrying short-wave infrared, thermal, and other sensors, these low-cost microsatellites will become valuable tools for surveying gas leaks and oil spills, monitoring geophysical disasters like earthquakes and volcanic eruptions, tracking air quality, predicting agricultural yield, and monitoring land cover and land use change as well as key parameters of water quality, including chlorophyll fluorescence, turbidity, total suspended solids, suspended sediment concentration, and surface temperature. These low-resolution, high-frequency multi-spectral data products will challenge traditional field measurements, becoming essential inputs to models predicting land use change, climate variability, and ecosystem and habitat change.

8.3. Integration of IoT with Satellite Data

3481 satle1141 27.12.2022 Satellite Imaging and Environmental Inference 1 8.3. Integration of IoT with Satellite Data Satellite data can be integrated with real-time and dynamic data, estimated by the Internet of Things. As IoT now includes a large number of widely dispersed sensors and is increasingly covering our planet, this integration can potentially provide novel data streams capable of vastly improving how, and at which scale, we use satellite data to estimate important environmental components, such as

those ascribed to the Earth system by the Land Climate and Hydrology themes, carbon monoxide, nitrogen dioxide, methane, ground level ozone, and other trace gases, and other constituents such as aerosols, particulate matter with diameters less than or equal to 2.5 μm or less than or equal to 10 μm, black carbon, soot, and atmospheric deposits of dissolved gases or particles. Integration of satellite data with IoT could also facilitate other nowcasting, forecasting, and modeling activities such as wildfire prediction, flood risk forecasting, coastal structure climate resilience modeling, storm surge modeling, and marine ecosystem management, beyond estimating concentrations and emissions of various trace gases. These contributions are the result of the collision between astronomy-based physics and more recent atmospheric process developments, creating opportunities for new partnerships to the benefit of both the Space and the IoT Worlds. In particular, satellites are advancing their frequency and spatial resolution while also investing on innovation around new sampling designs. In addition to the increase of the amount and types of data being obtained from terrestrial and aerial vehicles, this is opening new modeling avenues. Factorized statistical data-driven models developed for the inverse inference of IoT data are now being adapted for the use of data from satellite and air quality sensor networks.

9. Conclusion

In this work we have explored different aspects of physically consistent methods for modeling and interpreting satellite imaging data. The simultaneous modeling of acquisition and retrieval is shown to exist under complementary assumptions that we make on the radiative transfer and the imaging physics, mostly regarding the adopted approximations and constraints. We also show how to configure generic numerical solvers for the retrieval stage under a tree-structured prior. The combination of those two aspects allows for an efficient implementation of accurate principled algorithms for the inversion of imaging data. We validate the proposed methods with applications to define some environmental properties from different satellites and, more generally, we build toward addressing more ambiguous problems with the Monte Carlo approximation of the model. The achievements summarized in this work help posing an experimental basis for the analysis of richer datasets. In other words, they pave the way to more complex studies with increased physical richness and flexibility, such as fitting both the acquisition and the high-level parameters to perform the classification of human and natural classes, or modeling the time evolution of urban and habitat parameters within satellites to match climate studies. Those studies are important as they allow to bring recall and homogenize virtual observations from the different assets to increase the physical significance and quality of the common result. Finally, those analysis offer rich environmental insights at a much larger spatio-temporal resolution than currently available with other assets. Furthermore, the fast algorithm that we enabled should allow for the integration of those analysis within data analysis

pipelines, offering real-time or near real-time response, which itself could further increase the interest and applications of this technology.

References

[1] Ojo OL, Ajiboye Y, Afolalu AS, Morebise AT, Omoyajowo IM, Abe OE, Olumodimu O, Adeyeye DS, Agboola OM, Olawunmi O. AI Applications in Satellite Image Processing: Enhancing Earth Observation and Environmental Monitoring. In2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON) 2024 Nov 26 (pp. 1-5). IEEE.

[2] Diana L, Dini P. Review on hardware devices and software techniques enabling neural network inference onboard satellites. Remote Sensing. 2024 Oct 24;16(21):3957.

[3] Alotaibi E, Nassif N. Artificial intelligence in environmental monitoring: in-depth analysis. Discover Artificial Intelligence. 2024 Nov 18;4(1):84.

[4] Sloan S, Talkhani RR, Huang T, Engert J, Laurance WF. Mapping remote roads using artificial intelligence and satellite imagery. Remote Sensing. 2024 Feb 28;16(5):839.

[5] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.

[6] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.

[7] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.

[8] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.

[9] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.

[10] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.

[11] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.

[12] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.

[13] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.

[14] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.

[15] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.
[16] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.
[17] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.
[18] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.
[19] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.
[20] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.
[21] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.
[22] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.
[23] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.
[24] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.
[25] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.
[26] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.
[27] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.
[28] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.

[29] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.
[30] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[31] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[32] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[33] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[34] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[35] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[36] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[37] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.
[38] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.
[39] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.
[40] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.
[41] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.
[42] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.

[43] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.
[44] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 9: Interpretable Forecasting for Disaster Preparedness

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Interpretable Forecasting

There has been increasing interest in understanding the predictions of machine learning models for high-stakes applications in areas such as health, finance, and public services. Complex deep learning models, including neural networks and gradient boosted trees, have often been outperforming simpler forecasting approaches based on regression [1-3]. A common criticism against them is the difficulty of interpreting their predictions in practice. Since forecast models are often elements of integrated risk management systems, decision makers typically focus on the long-term policy implications of the future conditions these models are describing, rather than on the forecasted values themselves. Forecast interpretability thus refers to explaining the dynamic influences, e.g., fuzzily how future temperature and precipitation levels may on average affect the probability of flooding in a three-day time frame, rather than explaining the numeric prediction errors associated with prediction accuracy losses.

Interpretability is especially important in forecasting for disaster preparedness because improving performance on validation data, which is a natural goal in statistics, machine learning or operations research, does not guarantee better decision making during actual disaster events [2,4]. Also, many complex non-linear imputation methods used to fill in missing data can err in similar ways, which may be difficult to anticipate. A next step in addressing the above gaps is to develop an interpretable forecasting approach that focuses on hazard levels rather than predictions. In that spirit, this paper presents interpretable multi-hazard probabilistic forecasts, which are illustrated in the context of two practical problems, disaster loss management and flood risk monitoring. The rest of this section motivates a novel visualization technique, and concludes by outlining the two applied contexts [5-8].

2. Explainable Models for Early Warnings

An effective disaster response plan requires preparations in advance. Variable, hazard, and site selection are critical in judiciously making limited resources for disaster forecasting available in a way that leads decision makers and the impacted population to make timely decisions [6,9]. At the same time, early warning forecasting models necessitate interpretable feature and prediction relationships, to filter for forecasts that can be relied upon. Empirical explanations of how models make predictions and investigate model sensitivity to changes in features has the potential to instill people with an appropriate level of trust in the model, while informing them about aspects of reality that the model leverages, which may be non-intuitive. For a decision maker, understanding how model predictions change with changes in input features can assist them to anticipate the duration of the event and therefore prepare the appropriate amount of mitigation, mobilization, and response capabilities. Post-hoc explanations that provide interpretable relationships allow at-risk communities and stakeholders to gauge when to rely on a prediction from the model for the initiation of public policy action, and formulate their personal-level decisions if and when needed.

Three types of models exist: “white box” models are interpretable by design, while “black box” models rely on post-hoc techniques for explanation, such as feature importance scores, that can be challenging to leverage right. Also, the scope of the prediction-local and the task-oriented aspect of interpretability complicates the search for insightful explanations about model prediction for machine learning developers and domain experts alike [10-12]. Several types of explainable algorithms exist to investigate how and why models make predictions, and for which a variety of metrics have been proposed to contend with the heuristics involved in explaining model predictions locally or globally.

2.1. Importance of Explainability in Forecasting

Forecasts are seldom treated as absolute truths, but rather as statistical reconstructions of a process of interest. For numerous reasons, as change blindness and repeated exposure, forecasts are interpreted not only to predict future values, but also to explain the underlying reasons that generate the outcomes of a stochastic process, answering the question of why. In addition to generating accurate future predictions, researchers and practitioners attach significance to the elicitation of informative forecasts, also known as judgmental forecasts, to help explain, support or challenge fallacious beliefs, and to shed light onto the hidden mechanics of events [7,13-16]. These desirables are purposefully sought in various forecasting applications, from person based to mixed disease contagion, from extreme weather conditions to social development, and from sports to political events. In this work, we center around explainable forecasting

methods that uncover the hidden mechanics of the predictive process that generates the outcomes because we are particularly interested in using forecasts both as prediction and understanding instruments. The question of why is crucial in this type of forecasting because, within the domain of disaster preparedness, questions like why, how, when, or where will a disaster happen hound the expert's mind for broad preparation design requirements. In this domain, the apocalyptic sense of the question of when a disaster will happen is difficult to answer. It implies forecast durability, mitigation action design preparation for possible future disasters, and forecast design, production, execution and supervision for present moment disasters.

2.2. Types of Explainable Models

Two concepts underlie most explainable models: first, that the model itself can be inspected, that is, the mappings from input to prediction values can be made apparent to users; and second, that the problem addressed is simple enough that the associated model is "underfitted," i.e., the model is "too simple to do much wrong," and thus trust may be unconditionally conferred to its predictions [2,17-19]. In forecasting problems, we usually only have one interpretation of the model reported back to users, as they are not kept up to date with new meanings to the models. For short-term models, we implicitly trust the model, as their predictions are often quite accurate. For mid- to long-term newsworthiness models, their simplicity inspires trust.

The simplest interpretable models are based on historically blocking or subsetting input features. Blocked methods, such as "All earthquakes above magnitude x" or "All mass shootings in locations of type y," yield models that look at historical events and restrict predictions to those with matching parameter values. However, this identification-based rationale can be terribly misleading if it is based on non-explanatory values. Simple checks on the latest events help make better sense of predicted values in these cases, even though the predicted event is not set in stone.

Subsetting methods don't yield pinned events, but they do use a set of events whose trainings are allowed to inform predictions about other events. An example of a linear thresholded Gaussian model subset-based approach predicts an event using nearby historical events. Filtered models imposed penalizing filters on models to reduce space from a given source, in the service of emphasizing structural patterns, but they didn’t yet implement a learned to reduce stage.

2.3. Evaluation Metrics for Explainable Models

We review evaluation metrics for explainable models based on three perspectives: 1) Model assessment (i.e., semantic, visual fidelity), 2) User-based evaluation, and 3) Task-specific evaluation. These perspectives are informed in part by typical prediction

tasks for explainable models. The structure is a mix of theoretical and empirical analysis of model-assessment and user-based evaluation perspectives. More precisely, since we present mostly new user-based evaluations, certain aspects of model assessment are covered in more detail, but they tend to be rather generic and equally apply to more model-specific metrics. With task-specific evaluations, the more specific details blur the differences, but we tend to be more model-centric in this case as researchers are usually interested in specific answerable questions when developing those models.

Forecasting of time series for which explanation is required can involve point, probability, and quantile forecasts for several prediction horizons. Some approaches focus only on point forecasts which is commonly the case in the framework of explainable deep learning methods due to their mainly predictive task-based internal procedures [3,20-23]. Common probability and quantile forecast evaluation metrics for the actual forecast error at certain forecast horizons and training errors for deep learning models are the usual choice. Models can also achieve, however, task-specific goals so that they are indirectly evaluated regarding the use of their forecasted results. Lastly, although not actually forecast evaluation metrics per se, time series clustering or classification metrics can also help assess whether a certain explainable method could even serve its main use case, the interpretation of usually numerous and complex predicted structures of geophysical relevant quantities, to ensure reliable decision-making.

3. Multi-Agency Decision Support

Many agencies contribute in different ways at various stages of planning for and responding to disasters. For instance, extensive forecasting is conducted, including of weather phenomena like hurricanes that can generate the event of interest, down to hydrological forecasts that predict the river level and timing of flooding. Separate from this, agencies may work with others on flood models to determine how much flooding is expected for areas along the river and when, which are then considered alongside hurricane-driven storm surge flooding [9,24-26]. National resources may also be called on to be deployed when and where needed in the case of a notifiable event. Outside the United States, international responses to outbreaks of health-related issues may be called upon. In all cases, some common information, like what illness to prepare for, must be delivered and available from forecasting to operations.

This chapter presents an overview of information and data access requirements for these multi-agency responses. It is not intended as a comprehensive treatment, but rather to illustrate the different protocols involved in responding to health-related disasters through representative agencies involved in unique capacities. Providing

accurate predictions of disaster health impact timing, location, and severity that are easy for all response agencies to access, and that present similar information for various decision-making purposes, is foundational to successful collaborative disaster response. In many cases, forecasting and response coverage is at a national scale, and thus robust forecasts with clear accuracy and uncertainty information help target and prioritize local disaster response resources by decision makers during both the preparedness and operations phases.

3.1. Role of NOAA in Disaster Preparedness

The National Ocean and Atmospheric Administration is one of the first agencies to forecast, analyze, and visualize an imminent disaster's danger. They also inform the public about warnings, help evacuations, provide resource allocations, and deploy assets for response once disaster strikes [27-29]. Their activities include residential risk assessments, state and local contact coordination, interagency response preparations, and resource allocation—of people and equipment—to various threatened or damaged areas. They also coordinate throughout the disaster cycle, helping public entities to prepare for, respond to, recover from, and mitigate the impact of disasters.

Through this, their work facilitates the efforts made by local and state entities, such as voluntary evacuations and confirmation of building safety post-disaster. For health hazards, they collaborate with the health sector to minimize community harm. Communications are performed to schedule and receive reports, including staffing deployments, modeling evaluations, and dissemination processes regarding forecasts. Advisories are issued and updated periodically for general situations or specific events. Additionally, in conjunction with local-state governments, they display life-threatening event timelines to anticipate casualty impacts, leading to research prioritization and mission focus.

3.2. FEMA's Approach to Forecasting and Response

Forecasting addresses known unknowns that can be pinpointed in geographic space and often time. Some forecasts are made at regular intervals while others are unpredictable. What is unknown is when the event will happen but the season is known. For some events, track record magnitudes are unknown and the probability of an event occurring is unknown. What is clear is that the office for forecasts works closely with to minimize the impacts of events. Clear and timely forecasting help in reduction of disaster risk through programs aimed at preparedness, impact assistance, recovery assistance and community resilience.

How supports forecasting and information sharing. supports GIS support by working with state and local mapping agencies to identify high-risk regions, and ensuring

developed maps and models are used operationally for the mitigation, preparedness, response and recovery process. For decision makers in anticipation of a disaster, geospatial products provide critical localized information on potential social, economic and environmental impacts of the hazard [30-32]. With access to a great deal of GIS resources and contacts throughout all government levels, will work to assist creators and users of geospatial data for effective disaster response and recovery. Maps obtained from GIS will indicate which areas will get hit hardest by hazard dynamics, what populations will be affected by a disaster, and how customers may be reached for fast response. For disaster response, links and maintains contacts with recognized GIS organizations, offers vehicle and analytical resources, maintains GIS operations and infrastructure, and provides analytical tools such as disaster debrief protocols and impact evaluation survey instruments for response and recovery phases.

3.3. WHO's Guidelines for Health-Related Forecasting

The “Guide for the Development of Early Warning and Decision Support Systems for National and Community Disaster Management” serves as a crucial starting point for guidelines on health-related forecasting in a multi-agency setting because the intended audience consists of those who would be establishing the guidelines used by developing communities [9,33-35]. A summary of the guide and features that address how to produce better forecasts in a multi-agency system is provided. According to this guide, multi-hazard early warning systems deliver accurate and timely warnings that enable vulnerable people to take timely action to reduce disaster risk in a process led by disaster managers and not merely experts in hazard-related fields.

The guide has other key points regarding early warning systems. A communication and response plan is vital; the role of risk communication is discussed in-depth. Investing in disaster preparedness activities in advance can reduce the cost of relief operations and yield significant savings [36-38]. This takes the burden off the relief services and agencies, which will be better prepared to support the affected population. Activities such as contingency planning and stockpiling of key relief supplies usually require additional resources and commitment, but generally result in a quicker and more effective response. Early warning systems are designed to indicate the risk of a hazard reaching a specific place at a specific time. Hence, the role of other available data, such as the vulnerability and exposure data, needs to be considered.

Because the impacts of developing community disasters typically cascade downward into additional health and social issues, the description also discusses issues that affect the quality of non-health hazard forecasts and useful implications for model implementation. The forecasts are one of many decision support system services available.

3.4. Integrating Multi-Agency Data for Improved Outcomes

Decision-making is undertaken in many different agencies at selective phases of the disaster preparedness cycle. This includes not only mitigation, preparedness and recovery but also response phases, which serves to make these phases sputter through the year but also face acute needs at the actual hour of the event. Consequently these planning, management and action augment new communication technologies, which provide enormous new opportunities to strengthen intra governmental, inter governmental and governmental-nonprofit sector ties at many levels while at the same time straining this network in different prominent product by-passing, softening or immersion of government efforts [3,39-41]. This idea has been further extended for social media tools in the crisis response domain providing strong capabilities for organizations and managers that exploit these channels to enhance traditional and coordinated public communication.

Philadelphia, the fifth largest city in the United States, recently implemented the Philadelphia Fire Department's Firefly Initiative, which uses predictive analytics to identify neighborhood targeting at-risk and vulnerable individuals who may be prone to home-visit them for prevention of an inevitable fire which would result in intervention utilization over-response focused on all fires. The project establishes service thresholds, develops risk models, and examines factors behind breakthrough response. As part of this study we address ethical issues, commensurate with the Informed Client hopeful goal of the Philadelphia Department of Human Services, which requires that these prediction tools are not only useful but also usable and acceptable to multi agency teams serving community crisis prevention.

4. Case Studies

This chapter describes how interpretable forecasting methods (descriptive, predictive, and normative) have been used to prepare for disasters. The case studies share the common theme of being disaster-focused. Most are of dissimilar disasters, although the first two cases are based on predictive models for similar events: both are for wildfires, but one is an event occurrence model while the other is a fire-containment damage-avoidance model. The other three involve dissimilar disasters: floods, earthquakes, and epidemics. Despite the differences, there are some commonalities therein that are seen throughout the cases – e.g., the use of similar forecasting methods, problems in decision-making under uncertainty, the impact of climate change on disaster structure and models, and missing or incomplete ground-truthing records.

The final section discusses general lessons learned. The empirical predictions described in this chapter have been implemented by researchers in academia, government, and private industry along with the end-users in the respective forecasting

decision chain, such as emergency managers, planners, agencies, and companies. Both predictive and normative methods are emphasized. Notably missing, however, are any descriptive studies. The case studies and examples reflect the disparity between the theoretical importance of explanatory methods and their practical applications to date and suggest that there is much potential for expansion in the relative need for interpretable disaster forecasts.

4.1. Wildfire Alerts: Predictive Models in Action

Following the theoretical background presented in Section 2, we now illustrate practical applications and outcomes of interpretable machine learning and statistics in forecasting and modeling for risk mitigation on wildfire and flood. These two natural hazards trigger recurrent economic crises in a number of states and countries worldwide, and they are challenging threats to the global community in terms of disaster preparedness and consequent recovery.

Both research areas presented application examples of statistical and machine learning methodologies for different lead times. Thus, for short and medium lead forecasts of fire events, they proposed the combination of deep neural networks with rule-based classifiers of commonly discussed intermediate machine learning solutions such as bagging and boosting and classical state-of-the-art techniques. For long lead predictions (at least 6 months in advance), they proposed the application of simple deep neural networks such as regression and classification models with boosted and bagged trees for the prediction of fire occurrence and area burned in Canada, and bagging when forecasting area burned in Greece and the Italian regions of Calabria and Sicily.

Diverse explanatory input variables are considered for each proposed lack vision, including atmospheric variables, climate indices, and fire and lightning climatology. Also, in addition to predictive accuracy, their configured scenarios underscored some interesting economic advantages for the application of its solutions. For instance, in the specific case of the summertime decade-long fire events monitored in Greece, overestimated predictions were able to issue timely red (high-risk) alerts, more frequently leading to false positives without costly, additional expenses for ground operation.

4.2. Rainfall Predictions: Techniques and Challenges

Rainfall is a commonly cited example of forecastable meteorological hazards. Available forecasts from numerical weather prediction models exploit the principle of predictability, in that, by specifying their initial conditions with more skill, they can improve the predictability of weather for some upcoming time horizons. In contrast

with the task of elaborating a correct model to describe the dynamics and the potentially chaotic nature beyond the observed current state and its evolution, numerical weather prediction preemptively constrain the dynamics of atmospheric systems by performing a data assimilation step, letting the often minutely updated model output flows through all the computed domain adaptively adjust to the chaotic dynamics.

Despite the many advantages of being able to leverage numerical weather prediction model predictions and other weather observations, as well as the now widespread accessibility of hydrodynamic machine learning models, which are trained to learn the convolutions in their inputs needed to model decisions with transfer function convolutional neural networks, surrogate drought, albeit not flooding direct meteorological hazard variables, groundwater, and streamflow machine learning models have been shown to produce better prediction accuracy than first principles-based core hydrology models at fine spatial and temporal resolution prediction levels. Further, the potential of deep learning to learn latent hydrological predictive flows in streamflow/hydrology prediction without directly relying on use of numerical weather prediction model predictions or other co-variate weather observations has been abundantly and positively discussed, ever since the meteorological variable meta-category of controlled justifying explanatory variables of its input space was first introduced [36,42-43]. This is particularly true for non-permanent and permanent snowmelt and sediment transport forecasting, as well as very high and high resolution streamflow predictions.

4.3. Climate-Risk Scoring: A Comprehensive Analysis

Climate risk scores serve as a bridge between various predictive models and unstructured risk score drivers. These scores incorporate predictions from specific risk factors and accommodate likelihood-distribution uncertainty. This allows for a finer resolution of risk variations in windstorm, heatwave, and other regional climate change-driven hazards. Compared with machine learning-based and multi-grid resolution models, or parameters exhibiting natural resistances, or vulnerability-derived framework solutions, climate risk scores are a well-balanced compromise. They combine realism with an attractive cost-benefit ratio, and include both event severity and exposure as key risk factors. The result is a globally consistent, validated and event-exploring approach that is also highly adaptable to internal data-driven business requirements. Moreover, a shared, yet adaptable solution supports climate risk-category focussed product portfolios and clearer multi-branch division responsibilities. As such, climate risk scores may be employed to identify low-light risk locations for properties requiring special catastrophe covers, or developing product portfolios targeting individual climate risk domains. Furthermore, climate risk scoring

outputs can be used to align original spend and prioritization in the dedicated underwriting regions, and sensitivity reward-driven message align model outputs with internal sales processes. Climate risk scores can also be utilized as an optimal pre-consult phase-driven support for underwriting capabilities, for various natural movement cost assessments, or rank-based climate extreme-sensitive specified risk listing across global regions. Risk management is essential for understanding the inherent risk in a property, and for the establishment of strategies to deal with possibility and extent of consequences associated with such risks. Early identification of increased or changing levels of risk is an important part of managing climate risk throughout the life of a property.

4.4. Lessons Learned from Case Studies

Our case studies support the view that interpretable forecasting can close existing gaps between predicted and actual event outcomes for DP. Currently available forecasting tools tend to miss events, risk underestimation, or forecast irregularities rather than displacing gradual events. Climate-evolving events yield narrow lead times. While in some cases the automatic detection of forecastable events yields innocuous outcomes, in other case studies the provision of local tools boosts user trust, assists decision-making, and provides warnings of risk level-triggered preventive measures. Predictive skill does not always translate into economic value or improved preparation.

Despite the need to synchronize decisions across governmental and regional levels, our findings suggest that coherence of prediction harmonization is severely lacking in practice, heightening the risk of populational dishealth and hampering DP. We identify unsupervised nesting in spatial co-clustering as a useful predictor and warning of destructive climate events. Further, an analysis of the multitimescale leads to stable predicted probability density functions for cyclone- and tsunami-explaining covariates. Yet, the latter can still worsen predictivity. Predictive risk mapping via exclusive scanning window techniques can indirectly store information about critical thresholds, geo-localize it within a physical model, and assist in distinguishing between what may appear similar but is actually very different.

These case studies underscore the need to disentangle different predictive indicators before they can be used to jointly provide informative predictive distributions cross-timely and spatially. They argue against the popular wisdom that observation model tightness can substitute interpretation. The cyclone case study exemplifies that, despite prior reservations, econometry techniques can be useful predictive tools, especially for need-short term predictions until novel signals are lectured reliably.

5. Challenges in Interpretable Forecasting

Predictive modeling with novel machine learning tools faces a myriad of challenges in being adopted into real-world practice in procedural workflows that have historically used simple methods [6,9]. In addition to contextualizing a new model into established workflows with stakeholder input, a model also has to be meaningful and trusted to an audience that might not have data or programming expertise. This audience likely considers themselves the experts in the field and the messages that model predictions convey must be those which practitioners find useful as they pertain to the problem space reporting quantifiable and relevant results. Data science reports filled with new and exciting results alone are insufficient to generate interest. Therefore, we explore here some of the technical issues with statistical predictions, the need for tragedy informatics to mitigate loss of property and life, and the communication expectations of information users. There is a delicate balance between complicated statistical and data science assemblages that create exciting predictions and those that have a charge of duty to inform and enable harder socially responsible tasks such as disaster preparedness and potential mitigation.

The main paradox is that model quality is often at odds with the speed, interpretability, or frequency of updates that decision makers require. The selection of an appropriate forecasting method for a particular space and place, with the right temporal granularity, is not trivial. The tools from forecasting package solutions are often black-boxes built with out-of-date foundations. Transparent results are expected as regards error on fit period datasets and comparisons with equally simple but systematic other methods are rarely provided. In other areas of the social sciences, users expect transparency and a critical democratic dialogue surrounding model uncertainty about assumption violations, the appropriateness of goodness-of-fit tests, error metrics, and predictive evaluation. Predictive modeling in the data science spirit is not yet accountable to these fair ideals.

5.1. Data Quality and Availability Issues

Demand and supply data can suffer from problems like missing values, noise, resolution mismatch, multicollinearity, nonlinearity, and nonstationarity. Relatedly, demand data is typically attributed based on complex and proprietary data and algorithms given limited and noisy supply data. The effectiveness of the interpretable models and analyses, in this case, is limited. Clean data over long horizons is preferable. Given the socio-economic impacts a disaster can have on the demand and supply for some products, disaster events provide good out-of-sample test cases for interpretable models developed on clean data. In classical forecasting, or in supervised learning scenarios in general, using too much data can be as problematic as using too

little. The models devised are too complex, leading to overfitting. For interpretable forecasting, we find that the practitioners prefer interpretable methods as long as they are not substantially outperformed by their more complex competitors. In case local models are developed, avoid local outliers. The presence of such outliers helps the user make sense of local models, reducing but not eliminating complexity. While some researchers have overcome training and prediction delays, exploring most to all configuration and operational settings of complex military operations are significant hurdles, limiting the use of interpretable machine learning comparatively. Consider using simple historical models as benchmarks for comparison. The usage of such models is unlikely to suffer from data quality or availability issues.

5.2. Balancing Complexity and Interpretability

Throughout the history of statistics, models with a larger number of parameters or functional forms that allow for richer structures have been seen as more accurate for data heavy problems. For example, after the introduction of very flexible thin plate spline use in modeling, the interaction terms and polynomial power augmentations became less favorable options. However, even in forecasting, which is one of the main applications of statistics, the interpretability and explainability of prediction is important. Maybe unsurprisingly, the trend towards more complexity being more favorable has also been taking place. Deep learning, and certain recipes to fit very large models where induced and complexity ideas are tried in the data-driven prediction. However, the models are not very helpful when doing what the old patterns saw, as were depicted on the workshop dealing with departmental forecasting of budgets and enrollments, respectively, and the college-age population.

Furthermore, the expansion in the number of models being used has highlighted the fact that it is now infeasible to do the many checks on residuals, parameter estimates, and out-of-sample accuracy that should be routine for simpler parametric models. Therefore, practitioners have opted to use very complex but opaque models in certain areas like sports forecasting, deal-making, and potentially forming important aspects of recommendation algorithms in directing us to what is "popular." Theory seems to indicate that deeper models allow describe more structures if we indeed want to imitate the brain. On the other hand, humans are hard-wired to understand certain forms of prediction, and these may also be what our brains will think better about.

5.3. Stakeholder Engagement and Communication

Model interpretability does not guarantee insight. An interpretable model can be easy to understand yet yield unintended interpretations, leading to misunderstandings. Risk communication in the context of climate disasters goes beyond what occurs when an

interpretable model is misinterpreted to include communication of risks present in the model meant to inform parties involved in preparing for climate extremes, the model without advanced technical skills. Climate partner organizations can share how scientifically developed models inform decisions using standardized messaging. Their messages provide practical information regarding decision alerts, lead time, uncertainty, and sensible response options. These elements can usefully inform any climate or weather-based communication and decision statements and tools, helping to build public trust in models that guide responses. Model developers must familiarize themselves with model communications, combining their technical skills and those of partner organizations to create, deliver, and revise simple, intuitive statements based on interpretable model outputs. Such collaboration requires intention to harmonize message design and dissemination. It also requires commitment to co-creating communication training to develop both model familiarity and communication capacity in partner organizations, both at the model phase inception. Testing with intended audiences, including users and non-users, is important in understanding communication outcomes prior to dissemination.

6. Future Directions in Disaster Preparedness

This dissertation presents a new vision for decision-making based on forecasts. The work introduces a framework for interpretable multimodal predictions, in which human-centered approaches are taken to understand and explore the potential impact different sources of forecast uncertainty have on responses from disaster management agencies. Equipped with this new approach, disaster preparedness becomes collaborative, allowing the many agencies associated with these responses to share their expertise and provide feedback to complex model predictions. The future vision for disaster preparedness is made possible with advancements in machine learning algorithms that can converge with the challenging forecasting problems presented in the disasters, as well as the ability to leverage advances in computational techniques alongside important sources of real-time data pertaining to disasters. Our research creates new possibilities for actionable predictions, supporting critical decision-making at the most opportune timescales for disaster responses, and shifting research in disaster management towards data-driven approaches.

The vision set forth focuses on practical advancements in the application of disaster prediction science. In this section, we present 3 areas of potential advancement based on our key findings: (1) novel modeling developments enabled by advancements in new machine learning techniques; (2) the growing potential for real-time multimodal forecasting improvements through innovative uses of new data sources; and (3) collaborative frameworks that bridge monitoring efforts across multiple disciplines to help enable use-case-oriented prediction with a shared multi-agency objective. We

believe further efforts in these areas can forge a path towards our vision for disaster prediction, which enables better day-to-day management of disaster-related data.

6.1. Advancements in Machine Learning Techniques

Machine learning is an exciting research field where, within less than two decades, several meaningful results have been achieved in various domains, leading to major breakthroughs and state-of-the-art techniques that drive and follow the economy. For what concerns Deep Learning, the idea of learning useful feature representations of unstructured data led to state-of-the-art solutions in Computer Vision and Natural Language Processing. However, models for structured data are still at an early stage. On the one hand, traditional learning algorithms have been around for several decades, dealing from the beginning with challenges that many modern machine learning techniques are not capable of handling. On the other hand, a changing world and the big data era, through new sources of available information and better modeling techniques, provide the opportunity to exploit traditional algorithms to develop efficient, robust, and scalable solutions.

However, even if traditional algorithms still provide the backbone of most applied research, recently we have witnessed a growing interest in boosting traditional architectures with the flexibility of more sophisticated machine learning techniques. In this context, profound generative models have achieved state-of-the-art results in image restoration and generation but, to our knowledge, their combination with traditional world models is still unexplored. The role of generative models is important not only for discrete data, where classical optimization algorithms may require time-consuming tuning to solve the problems at hand, but also for structured data.

6.2. Potential for Real-Time Forecasting

Static forecasting is at risk for being outdated by the time of the disaster, particularly for settings where the data used for analysis are very old. Consider a hurricane approaching the US mainland, whose landfall it was predicted several days in advance. Open a historical file with cyclone conditions, as well as financial and infrastructure data. Suppose that you build a forecasting model today and let it internally estimate the coefficients that you observe on landfall day. Then you inspect the effect of different intensity values upon people's 911 call behavior, according to your forecasting model. You find that the intensity of the cyclone has a huge effect on those countries' people. The cyclone is intensifying. You announce your latest findings on the news, letting your warning come at the right time. This example shows that forecasting models which are built internally rather than externally have an advantage over such models

built with static analysis: Certain parameters need not be fixed for all time. Instead, the true effect varies with the stage of disaster evolution.

There is room for improvement. In our example, the model would have problems if the cyclone intensifies while some people are still evacuating. Thus phase-specific effects would ideally be modeled jointly rather than separately. We also need a proper linking of intensity to behavioral effect. However, these models are better than static forecasting since they account for the dynamics of behavioral change. This is very important because traditional static analysis is only able to capture one snapshot of a complex process in our examples when people are still allowing for an insightful comparison of behavior at the evacuation and the impact stage of disaster-related behavior. However, these studies need time to arrive at the public. Therefore, a real-time recommendation of parameter estimates is crucial in applied disaster social science research. In conclusion, there are promising approaches to find solutions to this need.

6.3. Collaborative Frameworks for Multi-Agency Efforts

Multi-agency efforts are generally present as part of disaster management, however, the need for collaboration and shared understanding has only been mentioned recently. At the management level, there are networks on disaster risk management and on prediction and forecasting in the global community. More locally, platforms exist for informal sharing of non-sensitive data. On the technical side, a key requirement for the adoption and successful usage of data science in disaster preparedness and other areas is an understanding both of the data behind the algorithms and of the models' mechanisms, dynamics and uncertainties. This has often been termed explainability or transparency, although in our opinion transparency is the more suitable term with respect to the mentioned conditions. Enhanced explainability is therefore a step in that direction, and many layers of this transparency are needed, of course with different levels of depth and breadth, depending on the audience engaged at each moment and occasion. These layers of explainability include presenting parameters and quantitative indicators of model performance, model design and model assumptions, as well as narratives on the background and predictions.

Collaboration and transparency between the agencies responsible for disaster management is absolutely a key aspect needed to optimize the outcomes of why machine learning models were designed and developed. Political decision making and negotiations are not at the level that science usually operates, but it is at this level and firmly established practices where cooperation and teaming is more difficult. Disaster preparedness machine learning modeling can point to the best practices and why, but that is only a small piece of the puzzle.

7. Conclusion

Interpretable and trustworthy forecasting is essential for public safety, especially for pandemic forecasting and disaster early warning systems. Unfortunately, classical forecasting methodologies do not focus on building explicit uncertainty models, and even the most advanced AI-driven forecasts are mere black-box solutions. Uncertainty quantification is important for time series regression analysis, but its application to point forecasts is limited both in their methods and explanations, not to mention the difficulty to quantify uncertainty for complex, heterogeneous models.

In this work, we introduced the task of interpretable forecasting and systematically studied temporal coherence — an effective strategy to unify probabilistic forecast distributions into aligned yet complex models — to explain existing probabilistic forecasting methods and increase forecast accuracy. Thanks to its temporal coherence property, we built a curator-friendly and interpretable methodology combining hierarchical fork sources, multi-level hidden states, and co-evolving latent variables, improving the accuracy of a number of forecasting problems and large-scale datasets affected by temporal patterns.

We hope that the ideas and methods presented in this work can serve as a step towards a deeper understanding of interpretable probability forecast distributions and their alignment with the real world, especially for time-varying risks like pandemics or disasters affected by the time of the year. There are still many open questions. How can we explain probabilistic forecasting beyond temporal coherence? How should we calibrate complex models forecasting different variable types? How can we achieve temporal coherence while dealing with various input and output forecast designs? What is the role of quantiles or consistent scoring rules in the explanation of probabilistic methods? These are some of the inquiries future research may want to tackle in the coming years.

References

[1] Alshayeb MJ, Hang HT, Shohan AA, Bindajam AA. Novel optimized deep learning algorithms and explainable artificial intelligence for storm surge susceptibility modeling and management in a flood-prone island. Natural Hazards. 2024 Apr;120(6):5099-128.

[2] Alshayeb MJ, Hang HT, Shohan AA, Bindajam AA. Novel optimized deep learning algorithms and explainable artificial intelligence for storm surge susceptibility modeling and management in a flood-prone island. Natural Hazards. 2024 Apr;120(6):5099-128.

[3] Diehr J, Ogunyiola A, Dada O. Artificial intelligence and machine learning-powered GIS for proactive disaster resilience in a changing climate. Annals of GIS. 2025 Apr 3;31(2):287-300.

[4] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.

[5] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.

[6] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.

[7] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.

[8] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.

[9] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.

[10] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.

[11] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.

[12] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.

[13] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.

[14] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.

[15] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.

[16] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.

[17] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.

[18] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.

[19] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.

[20] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.

[21] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.

[22] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.

[23] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.

[24] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.

[25] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.

[26] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.

[27] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.

[28] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.

[29] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.

[30] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.

[31] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.

[32] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.

[33] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.
[34] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.
[35] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.
[36] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[37] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[38] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[39] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[40] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[41] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[42] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[43] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 10: Scalable Architectures for Health and Climate Artificial Intelligence

Samit Shivadekar

University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction

In recent decades, artificial intelligence (AI) has been put to socio-scientific use in a variety of ways and on different fronts. In principle, there are few limits to its application for our main societal issues, such as energy or material consumption, urban infrastructure and production, ecosystems organization, mobilization, and other energy exchanges experienced as natural phenomena. However, up until today, the actual realizations of AI implementations and operations have all been limited exclusions, made public either through foundational philanthropy or unsolicited entrepreneurship, or kept private as specialized economies reliant on ubiquitous public resources [1-3]. At present and for the foreseeable future, an overwhelming share of modeling, creating, training, and deploying of AI functionalities requires disproportionately significant investment, competences, and knowledge. Consequently, AI still operates in the sense of an elite trade without replicable scalable organizational architectures. These circumstances obviously shape the expected results of using AI for interplay with “natural” or “social” systemic agencies responsible for most of the fatigue of contemporary human societies.

Furthermore, the goal of this essay is not only the display of possibilities that AI can offer in disclosing identifying, enrolling, and “using” material or energetic flows inside and outside, that is, in nature, our socioeconomic systems. It is also a demonstration of how the development of particular systemic understanding of the way systemicity dynamically selects flows trajectories may offer reusable knowledge in order to organize and finance the structuration of basic or advanced AI models and offer them

back to the nature–society flows in such a way so that the actual automation of accountants, decision-makers, and intelligent actuators be made circular, horizontal, "democratized" [2,4,5]. These basic realizations should be used to progressively lower the costs and risks of implementation, and thus, gradually increase the quantity and quality of AI organizational architectures scalable and repeatable at levels, and responsive to actual needs of the nature–society processes.

2. Cloud-native Design

The term cloud-native refers to an approach to digital service design and implementation that maximizes the properties of the public cloud as a platform for agile and resilient services that meet the needs of their users. A growing list of design principles and recommendations for cloud-native implementations are available from cloud service vendors and research organizations alike [6-8]. These provide valuable guidance not only for implementing services in a public cloud but also for the design of cloud-native applications that combine in-house data and development assets with remote cloud service components. They describe the principles of modularity and component-based design, the role of API contracts for interoperability, and the use of open standards for transport, messaging and identity management. They also highlight the importance of data management practices that abstract and protect the data assets, the need for resilient, self-healing fault management approaches, the value of stateless service components and the principle of automating as much as possible to minimize human error and effort.

The goal of these principles and recommendations is to help service owners achieve shorter development and deployment cycles, predictable scaling and elastic capacity, and efficient, low-bandwidth use of network communications, while also freeing service implementation teams from many of the burdens of daily and operational service management [9,10]. In a typical cloud-native implementation, solution components that are prone to failure are not only monitored, but actively recycle themselves for key failure types. Database calls are offloaded to asynchronous job-queuing services. Access control policies and authentication/authorization transitions through the service lifecycle are defined in understandable, literal policy files managed on distributed storage services. These capabilities allow teams to focus on service improvement rather than fire-fighting. Such a focus is key to enabling innovations in the accuracy and analytics of health and climate-related services that need to improve but often languish.

2.1. Overview of Cloud-native Principles

The term cloud-native describes a set of design principles that apply for workload intended to run in cloud infrastructure. Cloud-native workloads are designed to

effectively leverage the scale and elasticity of cloud environments, be dynamic and distributed, and leverage managed services offered by cloud platforms [11-13]. Cloud-native principles describe how to design and architect workload that is optimal for cloud environments, but are not an indication of what type of workloads can or cannot run in the cloud. While cloud-native principles were pioneered by the world of application design, the principles are also valid for many data-oriented workloads and pipelines.

Fundamental to cloud-native are scalability and elasticity. Cloud-native services are designed from the outset to be able to be scaled up and down massively – at a finer granularity than perhaps ever before, per geographic region or even per cluster for the case of machine learning – to absorb heavy utilization load during certain periods while still being cost-effective during the rest of the time. To take advantage of the scale of the cloud, cloud-native systems should be developed as microservices – collections of loosely-coupled, independently deployable services with one focus and line of business logic. Microservice architecture increase agility by decomposing monolithic builds into smaller components whose release cycles can be parallelized, decreasing time-to-market.

The third pillar of cloud-native applications is relying on managed services. Every cloud provider has a collection of services they offer per tenant. A premise of cloud-native design is to use these services and write glue code that makes components work together, not custom-coding every layer of the stack [2,14-17]. Examples of managed services include but are not limited to relational databases, serverless data processing, NoSQL data stores, Pub/Sub coordination, hosted Kubernetes services, and cloud resource orchestration.

2.2. Benefits for Health and Climate Applications

Data-driven and AI-first innovation in health and climate are the biggest hope to expand our understanding of the vast complexities in biological systems and ecosystems, and also to make the best use of all available data for decision making, for detection and tracking of outbreaks, for predicting future events, for optimizing resource allocation, and many other exciting possibilities. As data from various sources is increasing in volumes and complexity, making the best use of all possible data sources in increasing dimensions and complexities, to contextualize in time and location using appropriate methods, and inferring in the right way by deploying scalable and flexible probabilistic graph models are the key to gain useful insight, and making accurate and timely predictions. Bayesian Network and Probabilistic Graph Models provide us with the right instrument to combine various data sources, simple and complex relationships in a hierarchical framework, causal and correlative

dependencies available from other sources, the penalty on complexity and how complex on the space of sensible solutions in a smart way [9,18-21].

From a design and infrastructure point of view, Cloud-native AI enables close to real-time access to massive amounts of data in the cloud and provides distributed execution capability for many AI methods that are typically bottlenecked on limited CPU cores on a workstation/subcluster thereby facilitating near real-time learning or lowering turn-around time on research projects. Data freshness is one of the significant factors in achieving reliable inference, involved in various aspects of Health and Climate applications such as high-resolution time-series-based prediction models for climate, distributed training and inference in high-level models of epidemics, modeling for rare events in surveillance, forecasting and anomaly detection models for health. Scalable cloud-native architectures naturally explore cloud resources to learn from and compute validation, uncertainty, and decision thresholds thereby lowering turn-around time during the research phase or lowering the turnaround time for trust-building, help for improved support during the decision phase thereby improving confidence in using AI-driven insight for health and climate.

2.3. Challenges and Considerations

Deploying AI in the health and climate spaces is no small feat, as both areas involve processes and sensitivity to scale that shouldn't be overlooked. Moreover, these areas have unique risks which make deploying fragile prototypes risky, and force more complicated processes for transition from private cloud to public cloud. In the health space, private data governs the availability of a solution, while compliance and planning task specialization are critical to get right [22,23]. In climate, latencies and scale can be massively more involved, such as waiting for a hurricane to hit land, deploying sensor communication at scale, waiting several months for seasonal weather to iterate, etc. Both areas also involve multiple data partners or product integrators that imply greater responsibility management, balancing trust and public accessibility, and workflow configuration that often can't get planned out all at once.

Both domains naturally have data at multiple levels of granularity, requiring smarter cross-model training. While computer vision and speech tasks for human-environment-sensor modeling are more straightforward, abstract human behavior and policy allocation for training on patients or users requires producing or ingesting semi-structured data at a higher dimensional moment level. Bridging the single-enrollee training data gap with synthetic data becomes a crucial requirement, allowing models to predict the remaining effect on the people waiting to be modeled, and become increasingly effective over time. Above-norm faulty concentrations create more short-term, self-filling gaps, also reducing training validity time bins for inference systems

[24-26]. These models balance coverage and latency optimization differently at every profile phase, which can conflict across distinct goals of a federated training or batch cycle.

Localized but temporal and individualized phase-influence modeling can ameliorate these conflicts flexibly, energy and liquidity planning using pre-concise labels or agent data. Clinical patient implied decision intervals deduce and reduce possible input structures for consequential phase aggregation across locations. Mixing short- and long-term horizon modeling together could be a valuable research pathway, while competition could accelerate refinement of these models into robust product solutions with low price and user stroke risk.

3. Microservices Architecture

3.1. Defining Microservices Architecture refers to a high-level technical pattern for organizing software into distinct services. A scalable architecture for Health and Climate AI applications starts with the right architecture. These applications typically take a combination of dimensionality smaller than the full-dimensionality data, combining a large number of diverse sources about events and entities to assess phenomena and decisionmaking about the likelihood of conducting specific acts at specific times which may vary in scale spatially and/or temporally. They lend themselves particularly well to a microservices architecture, where pieces can be combined into a loosely coupled but coherent application program interface. Microservices refers both to the APIs and to the presentation of function and data by the piece, and also to the low level implementation of the pieces [27,28]. Health and Climate AI applications are likely to perform specific and global functions very differently, but they need to integrate seamlessly by taking advantage of knowledge representation.

3.2. Implementation Strategies The microservice APIs are usually exposed on the World Wide Web, and are implemented using a set of services maintained by what would be considered distinct program development teams within an organization. Standard practice is that each microservice would serve a specific task within the Health or Climate system, and that implemented as a “few lines of code” solutions calling down domain knowledge to access the task through an interface specially designed for that specific microservice. This task specific service provides. Small specialized code that can be reused by other domain teams within an organization, exposing the APIs that address the implementation of specific components of the Health and Climate tasks makes it possible.

3.1. Defining Microservices

Microservices decompose monolithic applications into smaller units, which progress and deploy independently from one another. This concept reinvigorates the way software is delivered as each Microservice provides modularized capabilities through well-defined APIs. Cloud computing and Serverless computing technologies are key enablers of this architectural approach. The advantages of Microservices architecture are well-known: independent deployability, technology heterogeneity, shared code repository, improved fault isolation, easy scaling, reduced impact of friction in provisioning, and reduced time between incremental releases [19,29-31]. Advantages such as time to market and high quality of independently deliverable Microservices compensate the developers suffering the cost of distributed environments, thus refocusing modern development towards the deliverable as a more important factor than the development environment.

Microservices are customarily designed following five guidelines. They include single responsibility, organized around business capabilities, independently deployable, loosely coupled, and owned by a small team. Single Responsibility requires that a Microservice is responsible for one action, which ensures high cohesion in the Microservice implementations and results in re-usability in other modern design techniques. The organization around business capabilities guideline states that Microservices need to map towards business capabilities for easy calls. The independent deployable guideline expects that a Microservice must be able to deploy independently of the others, thus requiring a robust structure that takes care of Microservices' interdependencies. The loosely coupled guideline requires that Microservices do not expose their internal structure and call another one through a well-defined API. Finally, the owner by small team guideline states that a Microservice must be owned by a small team that must ensure the Microservice has enough demand to be developed, evolved, and maintained, thus taking a business decision not solely a technology decision. For Microservices to be effective, these five guidelines must be taken seriously; otherwise, organizations will not reap the rewards offered by Microservices such as the invitation to focus on Small Is Better principle.

3.2. Implementation Strategies

Microservices are small, autonomously deployable services that communicate via lightweight protocols. Today, the problem involves architecting, creating and deploying technology "in the cloud" (which translates to low-cost infrastructure and the ability to quickly design, deploy, test and iterate technology) using APIs to enable microservices to connect together to solve a much broader problem than any one service could solve alone. These microservices focus on a single area of functionality

[32,33]. These microservices can be strung together using API technology to provide a specific solution. IT Services of the kind used by technology companies rely on cloud agency delivery infrastructure to deliver solutions that took years to develop in-house from big outsourcing companies. Enterprise medical systems like Electronic Medical Records (EMR) systems is perhaps the single biggest untapped market. These systems can cost in the order of tens of millions of dollars and effort – to code, test, and deploy. Technology partners should be looking at the microservices APIs that provide basic functions such as lab/pathology reporting, patient registration, encounter scheduling, clinical imaging, treatment order scheduling, treatment administration, pharmacy services, etc. These microservices exist in many medical organizations and need to be provided as solutions on the cloud to replace the custom developed solutions currently in use.

3.3. Case Studies in Health and Climate

It is informative to look at a couple of cutting-edge but non-trivial examples of architectures of the types we are proposing from the health and climate arenas not just for inspiration and context-setting but also for the varying strategies they employ. As already mentioned, there are mainly three challenges in deploying AI in the real world corresponding to the three D's: dimension, deployability, and data. To make earlier visioniers in deploying AI for social good with the adequate breadths and depths of their mission, we briefly discuss these projects. The first case study is MMHealth which has pioneered several deployed health solutions at multiple dimensions across the health value chain from prediction to prevention to diagnosis to intervention. The second case study, Clair which is also a deployed climate AI solution seeks to help realize the vision of 1.5° stability.

MMHealth is a health microservices system for a collaborative network of hospitals around the world to build, share, and inform the ML/AI-based "digital twins" of patients for timely, interpretable, and trustworthy actions [34-36]. It offers an evolving ecosystem with unique deployment strategies, involving privacy-centered real-world data collection with distributed data partner domain hospitals, clinical use-case-driven algorithm development and validation, model-release and monitoring as model-based recommendation-action engines, and telemonitoring infrastructure for risk-based tele-prevention, tele-diagnosis, tele-intervention, and tele-reporting.

4. Data Pipelines

The ability to work with AI-enabled technology solutions for health and climate hinges on the availability, accessibility, and use of the vast amounts of raw, cleaned, processed, label-enhanced, augmented, and integrated data that these AI-enabled technology systems can use. However, effective, reproducible, and scalable data

pipelines and other data solutions must tackle coordination and resource-related challenges as public, academic, and private data management capabilities and needs further develop and, as a result, increasingly complex multi-partner collaborative AI projects in health and climate become the norm rather than the exception. In talking with informatics and data engineers, developer scientists and researchers, and ML engineers, we identify key characteristics of good data pipelines: ease of access, support for reproducibility, data quality, support for diverse file locations, synchronization of heterogeneous datasets, support for multiple ML formats/data types, support for a variety of data runtime environments, multi-partner support, support for data exploration, governance, and compliance, and user support and documentation. These key characteristics are not universally required or prioritized for every type of project or every phase of a project involving AI-enabled technology, but they are feature priorities to consider when designing and selecting tools and solutions for data and its movement.

We will highlight various tools and technologies that emphasize these types of data pipelines and actions in our roadmap. For example, core pipelines may be looked at as presenting data in a standardized shape/schema that all downstream application users can rely on. An application that requires the cognitive and visual skills of a human should be run as close as possible to the actual pipeline which means localizing it to a small subset of data to move on-premise for the lowest bandwidth. Tasks that use considerable ML processing but require less human intervention or feedback may be run remotely.

4.1. Designing Efficient Data Pipelines

Data is the lifeblood of AI and ML, and in the case of Health and Climate, the quantities and complexities of data to validate and implement are enormous - small wonder the process for collecting, organizing, preparing, cleaning, and streaming data is one of the most specialized and challenging pipelines to build. The traditional process of manual data collection may take weeks or even months before software training can begin, with data streamed to an inference service after ingest - typically a background task. In contrast, the need for real-time validation means that a streamlined pipeline for continuous live data ingestion from connected medical devices for Health will improve accuracy and responsiveness. Perhaps most important is the price of cloud and edge services for these pipelines, which can add up immeasurably with very large data.

Old-fashioned brittle databases that rely on heavy use of SQL, JSON, on-disk flat files and schema requirements can hamper flexibility as well as sluggish I/O performance. Flexible cloud-native services such as cloud object stores, NoSQL and other

lightweight document and graph databases can assist in achieving low-cost, rapid development of solutions. However one of the leading emerging challenges to scalability and cost optimization of AI and ML in the enterprise is the emerging quantum of real-time metadata generated by cameras, sensors, and other connected devices for many domains. In particular, compliance and management of sensitive medical data for resource-constrained edge devices for wearable medical technology and remote patient monitoring will require advanced encryption techniques and packet structures. The need to minimize latency and monitor for compromised data bearing personally identifiable information will add to software development challenges.

4.2. Tools and Technologies

To create scalable architecture for AI and ML applications, suitable tools and technologies for data pipelines should be selected based on the required functionality of the data pipeline, data, and resources available to build the pipelines. These considerations will help define the scope of planning data pipelines. Generally, there are four major types of data pipelines with several specialized functions in each category according to modes [37-40]. These pipeline categories are batch processing pipelines, real-time processing pipelines, hybrid processing pipelines, and specialized pipelines. A batch processing pipeline's goal is to transfer large datasets from one place to another. A hybrid/passthrough processing pipeline performs parsing, transformations, or enrichment of records flowing through it and passes the resulting records along to other systems. Either of these pipeline types operates on a fixed schedule which means several hours or days of data are gathered before firing up the engine.

A real-time processing pipeline is responsible for continuously processing small slices of data that are then delivered to an application for consumption. This type of pipeline can make records available for consumption seconds after they are generated, although it typically takes several seconds to process the records. Specialized pipelines can use one or more different tasks and combine them to build a complex task into a pipeline, or can embed a third-party service and create an intermediary storage to create an integrated pipeline. Data pipelines can be created using templates or solutions.

A popular asynchronous data pipeline service is a powerful solution because it allows you to decouple your cloud service from applications, services, and dependencies. It lets you break up monolithic APIs and microservices. You can easily publish messages to a channel to make them available for other services to process asynchronously. Serverless solutions enable rapid iteration and scaling, and also a reliable queue for other services and servers. But remember, when designing data pipelines using this

service, it’s merely an async data pipeline and it doesn’t provide reliable persistence for your messages.

4.3. Real-time Data Processing

Real-time processing is a complex task that requires fast and reliable processing architecture and an intricate configuration which is tricky to set up and manage as it constantly needs to be calibrated and fixed. Below let's first dive into the complexity of real-time data processing. Real-time data processing is a unique combination of resource intensity with real-time performance requirements that focuses on the presentation and reporting of the data with guaranteed performance levels. For example, this application might take data from a logging pipeline or telemetry operation and report to a UI and communicate results including statistics and alerts. These applications often are a hybrid between batch and pure stream where data is allocated into batches and each discovery and result takes a near real-time processing effort [41-43]. These applications are demanding in that they need high throughput for ingestion as well as UI reporting and visualization presentation with low latency for both ingestion and results presentation. The data is consistent and high volume coming in from IoT, telemetry, and mobile applications with the demand for both analytical and alert response to the data events. These applications place more resource ability on the consumption side where the UI and analytics just need a constant data heartbeat to generate results based on rules and algorithms than the data ingest can be resource hungry. There are no addresses in the events coming into the data pipes; they arrive in streams and in bursts and focus on asset address information from the ingestion rules and reports for presentation. Latency tolerance timeframes are relied on pure ingestion application which is not micro-batch based.

5. Streaming Tensor Analytics

5.1. Introduction to Tensor Analytics

Tensors, as a generalization of matrices to higher orders, are multi-dimensional data structures that have attracted increased attention as we continue to navigate the data-rich era of scientific discovery. Tensors formalize the modeling of multi-relational data in AI/ML including temporal data such as videos, spatio-temporal data such as geoscience information and sensor measurements, and relational data such as knowledge graphs. Due to their intrinsic properties, tensors are widely used to accomplish the fundamental tasks of multi-relational data, including machine learning estimation of probabilities, and inference of missing entries and tags, parameterization, and structure discovery for latent variable models. Tensors serve in many applications of modern society including scientific discovery and advancement in space and earth

sciences, computational material sciences, life sciences, astrobiology, geophysical sciences, environmental monitoring and modeling, planetary science, and ocean life.

Machine learning and AI are now consistently leveraged in sensors that measure, quantify, and control the climates of Earth and beyond. With ever-increasing amounts of multi-relational data collected by Earth and planetary observing sensors, the data streams at unprecedented scale and diversity. Furthermore, the inference services that these data streams support are expected to facilitate seamless operation and broad adaptation of societies of Earth and beyond in synchrony with the climate. Machine learning and AI have successfully demonstrated their capacities to provide inexpensive intelligent services for many tasks expected of a scalable architecture to characterize, quantify, and model the climates of Earth and beyond. But there is also a huge gap between the results and expectations for inference accuracy and cost, which opens up immense research, technical, and business opportunities for scalable tensor analytics infrastructure that can boost the practical utility and deployment of intelligent inference services powered by tensor machine learning models.

5.2. Applications in Real-time Inference

We consider here tensor analytics for Its Cloud to support new services that characterize, quantify, and model the climates of Earth and beyond, with supervision from scientific and domain expertise. To meet the demands of diverse inferencing tasks, the planned tensor analytic services include supervised and semi-supervised inference, monitoring and ensemble forecast/batch prediction, scalable four-dimensional learning, embedding, and continual learning.

5.3. Performance Optimization Techniques

In the next several sub-sections, we detail performance optimization techniques for streaming tensor learning and inference. We first describe some tensor-space models and how they can accelerate tensor analytics.

5.1. Introduction to Tensor Analytics

Tensor computing is a promising computing architecture for handling real-time AI problems related to climate change and public health. In this work, we implement, optimize, and evaluate tensor-based algorithms. We build GPUs accelerated tensor networks to optimize inference in a variety of models ranging from CovNets to deep GNNs. The proposed methods achieve higher accuracy and faster inference than contemporary algorithms. We focus on real-time classification and detection problems, where multiple messages are ingested through streaming APIs, models are

continuously updated and improved through active learning techniques, and various solutions need to be tested in a time-efficient manner.

Laboratory analysis of health and climate data is limited by the available sampling period, but careful model design and pupil-in-the-eye sensor placements allow for localizing and detecting key events within a defined but limited period of time, such as volcano eruptions, natural disasters, or outbreak of viruses. Solutions to these events have the capacity to inform larger decision making processes, but they need to be accessible and performed in real-time. Multi-dimensional data collections are being used and augmented with continuously streaming data from geo-sensors. The data needs to be analyzed for model selection in complex AI pipelines linking remote sensor observations and laboratory analysis, while also identifying the need for new laboratory data collections. Tensor decomposition-based methods have been proposed for these problems.

The previous work focused on either separate traditional matrix-based classification methods or high-level architecture design for tensors or on lossless data compression for GPU architectures. We extend tensor analytical methods to address multi-dimensional data analysis for event detection and localization, task assignment, and model selection for climate AI powering models deployed as online proxies or along online inference pipelines in healthcare.

5.2. Applications in Real-time Inference

Streaming tensor analytics platforms target real-time situations where tensors become available incrementally at high data rates and decisions must be taken as quickly as possible. Many such applications arise in domains where sensors continuously collect, accumulate and transmit high-dimensional data at very high speeds. Some prevalent examples include surveillance systems that continuously process streams of images and videos collected for targets with a potentially infinite number of spatial and attribute attributes, weather monitoring systems that track tornadoes, cyclones, storms, tsunami, floods and landslides through data from environmental sensors and remote sensing satellites, networks of smart grids and smart meters that monitor power outages and theft, nanomedicine systems that procure signals from nanobiosensors for monitoring the time-varying levels of viruses, proteins, hormones, etc. for different public and personalized healthcare applications, mobile crowd sensing and machine learning-assisted applications for a variety of city management services and for several personalized environmental monitoring services; and healthcare systems that collect personal health data through wearables, secure cloud environments and data service infrastructure to provide personalized care.

Data streams in all of these applications produce rich yet dynamic information of interest. However, either due to the continuously exploding volumes of visual, environmental, healthcare or smart city data produced or due to the limited budgets and collection lifetimes, just handling this information to develop scalable, accurate and efficient inference methods is a challenging task. In traditional tensor inference literature, as well as in most of its non-real-time counterparts, models are typically either absent or are simply built offline by pooling the available tensor observations over the entire data collection window. However, as any real-time monitor would agree, building a model merely on occasional data samples is an almost impossible task. In addition to the glaring analysis shortcomings, in utilizing such models for inference, one has to frequently reload and reread huge amounts of occasionally available data to make most transformation-related decisions.

5.3. Performance Optimization Techniques

We introduce a computation model for achieving perfect throughput in streaming tensor analytics. Given the building block of tensor analytics, we first extend it to a sequence of tensor-structured input data to enable computation over a stream of arbitrary length, given by Z = f(X). Streaming tensor analytics also supports complex data dependencies between the output tensor Z and the input tensor sequence X. The model contains a set of unit tensor element operators, corresponding to each input-output element pair in the equation. The operators apply the element-wise mapping zi = fi(X) for streaming tensor analytics and prepare its output from input tensors, buffered temporally. These operators can be executed on a streaming processor or a cluster of streaming devices, with multiple input-output pairs in parallel, to obtain the output tensor sequence {Z}[t−wmax,t], where t is the current sniff time in X, and wmax is the maximum output buffering delay. Our emphasis is on optimizing throughput during streaming inference and not for the entire async analytic pipeline model.

Despite all exciting research into the analytical complexity, fundamental limits of approximation, task graphs, and inter-task communication latencies, in order to introduce practical, efficient methods and models for realizing high throughput streaming inference for tensor analytics, we have to dig deeper and analyze the problems of bottlenecks and parallelization in the pipeline model itself. In particular, while specialized inference pipelines can exploit task parallelism, inefficient memory access patterns and poor input-output locality during inference are known to be major bottlenecks in achieving high throughput at scale. Tensor representations compactly encode the memory transfer volumes, latencies, and virtual memory access patterns for mapping data streams into inference pipelines. Exploiting the tensor data model

prescriptively, then enables designs that improve the aforementioned locality issues and parallelism bottlenecks.

6. Edge-AI for Real-time Inference

As introduced before, a major complication in the deployment of AI-based models for health and climate is the fact that many of them cannot deliver inference in real time or that inference has limited scalability. This is especially the case when models are deployed on large datasets such as those based on sensor data measurements over large geographical areas or those that track changes over time. For these reasons, it is important to utilize or develop Edge-Centric AI models, which combine the benefits of Edge Computing with AI in order to achieve scalable and fast edge AI inference. For example, for brain MRIs or other complex images, image classification or detected-object modeling would involve heavy computational costs since it might be infeasible to deploy the trained models remotely for each individual patient's opportunity request. Especially, Edge-AI for health-related AI inference is needed for situations involving natural disasters, military interventions, or other emergencies whenever time-sensitive risks are prevalent.

Integrating Edge Computing with AI Models is an essential part of facilitating these local inference procedures. Edge AI enables local data processing, local intelligent inference, fast communication, more added-value services, and secure central control. Different types of data can be efficiently preprocessed near the data source before being sent to the Cloud, central office or HQ. Preprocessing near the Edge can greatly reduce the power consumption and cost.

Because of the characteristics of climate data sensing, collecting, and transmitting, and the volume of climate data collected, Edge-Centric AI for climate research is indeed content-relevant. Phenomena such as tornadoes, hurricanes, floods and droughts involve extreme events that are expected to become more severe and frequent and yet are still often poorly simulated. Research on the optimal placement of Edge AI-Centric architecture under a hierarchical setting can provide important insights into enhancing the predicted reliability, robustness, security and accuracy of climate change monitoring and climate crisis management.

6.1. Concepts of Edge Computing

The edge, or more precisely the Edge-AI concept, aims at optimizing this last mile of the sensing-processing actuation AI data pipeline, which is associated to an edge device. The goal of the Edge-AI concept is to execute parts of AI inference at the edge of the network in real-time, and to execute the remaining complex parts of AI at the cloud in batch mode. The minimal model size for AI is the de facto Edge-AI

configuration, where the entire reason to believe layer executes at the edge in real-time and the complex AI model executes at the cloud in the offline batch AI inference. The edge must first and foremost have real-time autonomous actuation capabilities since the goal is to develop a deterrent cyber-attack architecture.

The operational distinction of Edge-AI is that the inference needs to occur in real-time at the edge, which in turn requires a minimal footprint model in the reason-to-believe and the actuation layers. The core idea of Edge-AI is to identify the layers of AI inference related to the R2B and the real-time actuations at the edge, and then deploy these layers at the edge for real-time actuation. All non-real-time plateaus of the inference pipeline can be executed at the cloud for higher efficiency and accuracy. The action and R2B layers of deep networks corresponding to the outputs of these layers should be much smaller than the original network, and can deploy models with an accuracy nearly close to that of the original larger network. These smaller networks or parts of networks can be constructed using a wide range of techniques. Various techniques have been proposed to quantize, compress, prune, and distill a large model.

6.2. Integration with AI Models

AI can automatically find synergies of the relevant dimensions for specific applications. For a task defined by an objective and a set of training examples, the architecture defines how individual examples are transformed iteratively to converge towards the optimum. Multitask and multi-input architectures allow AI to discover low-complexity representations by leveraging similarities between tasks, inputs, or both. Coordinated layers constrict representations, by following the object identity signalled by an input or task. Very large models trained with much larger data on non-unique related tasks discover such integrated representations. Specifically for Edge-AI serving time-critical applications, narrow and deep architectures allow much quicker inference. For real-time health surveillance, events have characteristic time scales, related to disease dynamics and the way the health signal is collected, and large changes in a class at a pre-defined frequency identifier-an event; at the particular temporal resolution of the modality, are detected. Events, anomalies, or discontinuities are efficiently resolved by difference and derivative calculations.

Edge devices are computation-constrained, often battery-operated. Models on device must be as small as possible, discretized, and neuromorphic, performing the equivalent of parallel inner products of feature vectors followed by conditionally actuated activations. Mobile applications must aggregate over population groups, using localized models on device, preprocessing raw data to extract shareable lower-dimensional embeddings, and uploading to a central Edge service where they are pooled or aggregated to update model parameters. Edge-in-the-cloud does not incur

latency for model training. For real-time inference to serve critical applications, models can be trained with transfer learning on new examples, agnostic to modalities. For Edge-AI supporting mobile access, the cloud must use swappable local cell closing with multi-sensor satellite and drone modalities that upload geo-tagged lower-dimensional embeddings for coalescing training.

6.3. Use Cases in Health and Climate

The COVID-19 pandemic triggered the need to equip citizens with warning tools based on reliable, real-time, high-frequency datasets. The use of wearable health devices is no longer limited to lifestyle and fitness tracking but has shifted toward monitoring individuals' critical health conditions, from sickness to recovery. During the pandemic, devices like the oximeter jumped into public use and monitored patients' blood oxygen levels. Smartwatches and fitness trackers began to display new information, such as abnormal body temperature, heart rate variance, and blood oximetry. Researchers opened discussions about the use of Artificial Intelligence on health wearables data. Distancing and lockdowns prevented physical contacts posing barriers for professional physical examinations necessary for regular disease monitoring. Citizens became more concerned about their long-term conditions and were always looking for non-invasive ways to contact physicians for advice. AI-Based applications started to help in disease detection through constant telemonitoring and remote reporting. This sort of cost-effective preventive measure offered by new-generation devices was covered by health insurance.

The search for a direct correlation between premature deaths worldwide and subsequent historical causes linked to climate change is no longer carried out only by specialized agencies. The world is now more aware and asking for regular reports input. New technologies show AI will play a role and already work to support, simulate, and improve findings. Researchers and policy-makers are aware of the synergy between Health and Climate and are joining forces to develop explorative models to obtain simulated-impact-based future scenarios.

7. Secure Data Sharing

Securely sharing sensitive data is a fundamental concern where the delivery of digital consent is a matter of life and death, as is the case in health and climate. Further, the owners of the data often have a clearer idea of how that data should be used for mutual benefit. In these cases, having a secured and trusted environment is essential for both the owner and the user. An organization running ethical climate studies may need data from an emission detection model run by a user on the edge. Secure outputs in a version-controlled manner would provide valuable information to both parties. Unfortunately, there exists no secure solution for these use cases today. Existing

options include either public exposure of the potentially-sensitive output or requiring the user to trust the requesting data study organization with their outputs.

Further, with stricter compliance requirements, such as existing compliance and future regulations suggested by organizations, increased trust among the entities in a collaborative architecture with end-to-end security will be the difference between a successful architecture and an ethical failure. Other methods may be unable to resolve these privacy requirements due to other constraints, such as excessive cost or lack of regulatory compliance. A real-world requirement for compliance is being able to attribute any breaches in security since they might reveal sensitive information, such as the ethnicity or financial strength of a patient. There has yet to be a model proposed that allows on-device AI deployments while also embedded with ethical solutions for secure data sharing.

7.1. Importance of Data Security

A notable tradeoff between achieving the ambition outlined and the increasing negligence of user privacy is the question of the importance of data security in a shared architecture. Healthcare and the environment are two sensitive domains that generate a large amount of publicly available data. Additionally, these two domains present a high potential for unethical and harmful model inversion attacks. Model inversion attacks work by figuring out what individual subjects are like using shared model parameters or, if a sufficient number of explicit outputs are available to craft a local model, reverse-engineering a local model that predicts an individual subject's outputs. Furthermore, such an attack becomes easier if the attacker is familiar with the population.

In healthcare, user subjects are individuals, and each subject has a personally identifiable history linked to potentially sensitive medical consultations and treatments. In climate, user subjects are social agents, and such agents are sometimes linked to identifiable businesses known to produce high amounts of carbon emissions or deforestation. While sharing non-sensitive models trained on aggregate data is perfectly acceptable, applying architecture sharing to sensitive healthcare and environmental data is an abstraction that conflicts with the ethical and legal principles of the two domains. Principles that deeply govern policy and decision-making in these domains are based on two factors: risk and perceived risk.

7.2. Methods of Secure Data Sharing

This section details methods of secure data sharing that are generally applicable to health and climate domains. We encourage investigators to seek collaborators with expertise in data security and legal guidance in developing secure data sharing

procedures. Research protocols typically include highly sensitive private data in large quantities, best managed by a small number of specific institutions, given limited data sharing incentives. Limited and targeted data sharing can ensure that high-risk data are stored and curated only in specific organizations, including domain-specific repositories for genetic research, secure servers, resource-intensive or frequently-updated datasets on high-access-need abstracts. Cloud-based services with strong security do provide rapid access and can enable investigators to build scenarios.

Data from private organizational servers can be shared to external collaborators in either of two different approaches that protect data: Identify-external control/agency-based solutions keep granular information about data identifiers and possible actionable data housed together in secure local servers at external collaborator institutions that guarantee high accountability. Support-shared identifier-based solutions create a structural mapping between distant identify changes in database and development of a controlled set of supported mapping identifiers that could never make the shared data usable by third parties. An estimate of the identifiers given vulnerabilities for many data use cases using encryption, secure multiparty computation, or differential data privacy could provide useful guidelines for queries. Aggregated data groups by the pillar identifier, temporally delocalized access could also be appropriate for either solution.

7.3. Regulatory Compliance and Ethics

Beyond security, legal compliance and ethics of data sharing are major concerns. First, data holders have to check whether their data creation/regeneration process is compliant with relevant regulations. If the answer is no, the data cannot be shared regardless whether a secure solution is in place because the organization may be subject to severe penalties. These violations stem from legally established principles on which several laws rely. Regulations are not only about data transfer to third-party organizations, but also about the purpose of the data collection and how the subjects of the data are informed. In particular, principles of lawfulness, fairness, transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity, and confidentiality must hold when collecting personalized data. Furthermore, there is a strict definition of patient data, as well as a specific protocol of how to handle it. Additionally, many countries have recently promoted new laws directed to the ethical use of AI technologies. AI systems must avoid bias and discrimination against historically marginalized groups, offer appropriate recourse measures when things go wrong, and show transparency and explainability.

Second, even if the compliance check is passed, organizations holding sensitive data may still be reluctant to share/publish it, as well as associated features not protected by

regulations. Secure solutions implementing differential privacy may use parameters or be configured in a poorly designed way. Releasing more criteria about differential privacy mechanisms would reassure stakeholders that the proposed methods are adequately privacy-safe. Addressing ethical concerns on data sharing is a growing topic in machine learning. Broadly, two methods have been discussed: improved methods that address the shortcomings of traditional approaches and solutions that mitigate historical biases.

8. Federated Learning

Yet, multiple analytical purposes demand joint access to health or climate data on the level of single individuals, such as personalizing interactive systems or studying health or climate disparities. In this case, federated learning could enable joint model training without exposing sensitive data to each other. Federated learning was developed in the context of deep learning and enables a set of parties to train a centralized model collectively without accessing each other's picture datasets. Each party trains the model on its own data and sends the trained model parameters to a central server, which aggregates these updates and uses them to improve the centralized model. The central server orchestrates the data exchange but never has access to the data. Federated learning is based on the observation that AI models can use a useful amount of knowledge after being trained through many iterations on huge datasets without having access to the dataset in a standard way. This solution allows parties who do not trust each other with their local data to still benefit from joint model training and thus larger training datasets.

In health, federated learning has been proposed for problems such as estimating the risk of brain aneurysms or predicting the severity of COVID-19. In climate, its use for coupling Earth system models has been explored. Though very promising from an ethical point of view, federated learning is still comparatively immature and participation is cumbersome. Second, federated learning is a challenging solution that requires a broad research basis and the infrastructures to implement it in certain niches. In particular, it is hard to optimize a centralized model through the contributions of local models that were not trained more than on five epochs. This is especially true if those datasets differ starkly in size and nature. We will not be able to do federated learning without overhead and some knowledge of the model to be trained.

8.1. Overview of Federated Learning

Federated Learning is a novel approach to machine learning that enables multiple individual users to train a common model collaboratively, whilst minimizing the sharing of sensitive and private data. Centralized machine learning has shown impressive success across numerous domains, yet data-driven health and climate

science applications often require individual user data sharing to achieve model generalization and utility. This is especially critical in highly heterogeneous user settings, where individuals avail very limited amounts of data with potentially idiosyncratic patterns. However, sharing private user data can cause major privacy concerns or heavy regulatory constraints, especially in health domains, where there are strict legal guidelines on sharing patient data, and in climate science, where the original data may belong to individual users. Further, centralized methods for model training present additional barriers, especially in health domains. Sensitive data, such as genomic data, can be gargantuan in size. Storing that data in a central repository, even temporarily, is both costly and time-consuming.

Federated Learning addresses these problems. On the surface, it is a distributed optimization solution, wherein model parameters are jointly optimized by iteratively exchanging updates between a central server and multiple distributed clients. Clients running local copies of the global model use their own data to update model parameters, and send these updates to a central server. The server aggregates updates to get a better version of the model parameters, and communicates them back to clients to repeat the process. FedAvg, the first federated learning method, enables efficient training on distributed client data by using a synchronous variant of a simple algorithm. Clients run local stochastic gradient descent for multiple epochs before sending their updates to the server, thereby reducing communication cost. Data heterogeneity across clients leads to many challenges in implementing FedAvg effectively. These challenges have motivated a host of subsequent methods. For instance, adaptive methods take data heterogeneity into account when optimizing, enabling faster convergence. Other solutions adopt new frameworks for reducing downtime during the optimization, such as allowing asynchronous methods.

8.2. Benefits for Health and Climate AI

Federated learning allows for training powerful but complex models on sensitive data scattered across institutions without requiring organizations to pool their data, which could otherwise violate data sharing agreements or data privacy regulations. There are additional potential advantages specific to the realms of health and climate. Federated learning allows for a much lower barrier to developing AI tools in sensitive fields and have those tools be leveraged by many organizations while potentially enabling more fair solutions that are more aware of marginalized groups. Federated learning allows for the scalable development of generalizable AI methods that can produce better results for many, while intentionally avoiding the biases often hardcoded in non-federated products from consideration. These solutions may additionally require less, if any, data sharing.

Due to both data privacy and data sharing concerns, there is often a lack of large global datasets available for training the powerful AI tools of today. Achieving good model performance for these fields typically requires fine-tuning the model that was trained on the largest dataset available. However, the risk of overfitting to a single location is very high in these fields. Federated learning allows for a large initial training on diverse, though sensitive datasets to ensure better performance without the dangers of overfitting with fine-tuning on only a single dataset. In fields like health and climate where the repercussions of failure are high, utilizing techniques like federated or multi-task learning are essential for developing truly generalizable tools.

8.3. Implementation Challenges

Although FL has an inexpensive communication cost and delivers a more personalized and higher-utility AI model than traditional approaches, the implementation of FL has several critical challenges. First, FL introduces additional non-IID data and system challenges. Non-IID data refers to the straggler issue, where only a few clients may generate model updates based on a small number of training samples. Existing studies have shown that the performance of federated-mask training relies heavily on the selection of the learning rate schedule. The difficulty of the worker-specific optimization hazards will cause the divergence of the global model. Such a divergence phenomenon severely degrades model performances compared to traditional centralized learning. The server may discard the model update sent by a single client if it infers that the uploaded gradient is contributed by a few outliers, which could increase the generation of outliers from untrusted clients and also other non-IID problems. The non-IID problem is further complicated when the implemented AI model has different memory and computation footprints. Another critical challenge is that FL suffers from a lack of monitoring mechanism for edge and server devices.

Considering the privacy concern of end-users, confidentiality and integrity assurance of the global AI model in FL research is intrinsic since model update aggregation and sharing have a lot of privacy risks. As the parameters of a model may capture sensitive information that can be utilized by attackers to expose the private training data, the aggregated model may leak sensitive information about all clients. These privacy problems may cause stigma and social distrust among the users who may have sensitive data in the training set. The stigma and distrust about critical data can make the model inaccurate in sensitive task areas and, in turn, the service provided by the AI model may not be reliable. The non-IID data and stigma problems are particularly important in health domain AI solutions, where the data have privacy and trust issues. Data and model integrity assurance is essential to avoid adversarial examples in deep networks. If a malicious user succeeds in proposing an adversarial example that is mis-

classified, the performance of that AI model in real-world operations may degrade and, accordingly, may generate a service threat.

9. Integration of Technologies

This section describes the available and planned n-tier stacks, exposing the functional components needed to easily integrate and extend health and climate AI systems. We highlight the interaction pipelines using cloud-native and serverless technologies and microservices, exploiting caching and Pub-Sub models to share among systems intermediate and near-real time outcomes from core AI pipelines. These add-up to the batch pipelines triggering compressed batch processing and long-term storage of AI outcomes typically based on data lakes aligned with cloud Processing and Analytics services. Such n-tier stacks are designed to optimise the integration and security of third-party modules and shared infrastructure, co-location of core services and data, as well as near-real time batch-triggering and in-batch notifications required to orchestrate and monitor the parallel processing of fragile multi-source and unforeseen geolocal mappings.

We also introduce the principles and available tools for SOA-based interoperability. It is implemented between activable components providing REST APIs and/or exposing Pub-Sub channels, via a front API gateway and event routers modules. Thus, it advertises APIs' location and components notifying of events triggering third-party services activations, open to external microservices and modules sitting behind their private channels and gateways. Such automations are coordinated through a SOA interface defined meta model and catalogue. It allows model creators to expose catalogue items' templates, enforced by defined hooks and security constraints. It is used by Data Administrators to schedule flows, which has a light footprint. Thus, we take advantage of previously scheduled data caches able to reuse channel paths to be invoked upon changes in inputs published by third parties.

9.1. Combining Cloud-native, Microservices, and Data Pipelines

Developing a scalable architecture for health and climate AI is about more than simply selecting a technology stack. At each layer of a software stack, different technologies interact together to produce real-world systems. When we develop a specification of a computational system, we may be working at any of a number of levels and then we may be able to descend down to lower levels or we may need to map across domains. For instance, we may specify a pipeline processing satellite images that we then implement on a cloud native microservice stack.

We also need to carefully consider the shapes of the data as they flow through the various layers of the technology stack. They may flow through the various stages of a

pipeline for media processing as transitory artifacts, stored in different forms in a database, or batched on disk or in memory for processing by analytics functions. In our work, this various data shapes usually lie at the interfaces and the key points of integration between technologies. For example, the microservice that is installed onto a clustering of serverless functions may be invoked by request-response or other protocols. Stored data may be managed by cloud services or have a more conventional access method. These services and functions may be automatically controlled by orchestration tools or called up from a standard programming language. Functions on cloud storage may be accessed by batch and streaming services.

9.2. Interoperability Between Systems

Interoperability is a core principle of scalable architecture used in design, documentation, and deployment of software applications. It is an essential part of scalable architecture as it defines how two independent applications can work together to share data and tasks. An example of allowing two applications to share and communicate with one another with defined interface and contract is the use of APIs or application programming interfaces. The API ecosystem has become complex with different standards of API represented by various interfacing applications, in addition to a myriad of standards of document-based messaging through different formats or protocol-defined messaging. Relative advantages and disadvantages of each of the available choices of implementing API are trade-offs between bandwidth efficiency, latency, ease of use, and discoverability for request-response lifecycle of data flow.

Existing software tools and libraries have been created to offer these different types of available API in boxes. Either implemented by others and made open-source or released as commercial offerings. Furthermore, there are software tools that offer API wrappers for ease of use for developers to quickly encapsulate any underlying data flow faster with less hand-coding, thus speeding up the development of vertical applications. Additionally, there are available tools from middleware developers for managing and monitoring all types of API services for developer services. These tools also include authorization and access control services. API gateways allow internal and external APIs to be listed, called, and permission-checked for both types of services.

9.3. Future Trends and Innovations

Hybrid cloud and on-prem systems will continue to become more common, particularly in health and climate where we’ve got long data investigation times but potentially need the processing of the cloud for very short windows of time. Thereby enabling reduced friction and improved collaboration between researchers, scientists, and the tech platforms, we will see highly specialized role-based data sets with limited

compute support are made available and rack space for charging for both access to the resource and data will be used by larger organizations to seamlessly share their specialized data under very limited conditions. We are starting to see the impact of very large foundational models trained with compute at the discretion of scientists who sees value in creating enormous integrations of not only health data but also other risk related areas – such as genomic, and other anthropological insights – can be mixed with climate modeling data. With inter-operable tools spanning all of those specialized domains, recent research demonstrates that this sort of foundational work can lead to vastly improved new generation tools for both data analysis and risks modeling.

Highly structured resources will be created to be built upon while remaining locked down to brass tax so that massive scrutiny can be put into those foundational resources. The next big jump in scale won't be from the use of the current resources available but from the researchers able to authorize the clear use of the wide variety of tools and technologies needed to solve these massively multi-dimensional yet converging and distinct challenges. We are increasingly working with consortiums of local health policymakers with wide insight into both health trends and distinct health indicators for various groups who use their authoritative voice to both clarify the parallel but super-distinct work that each scientist is doing – often with disparate groups, data access, and research goals – can be peeled off so that all the work is transparent yet distinct.

10. Case Studies

To illustrate the working concepts of scalable architectures for hybrid digital-twin AI, we summarize a few projects and their outcomes. The novel reusable infrastructure building-bits described above were applied during these implementations, allowing us to avoid falling into common architectural challenges and scalability traps. Notably, there are no other examples of similar multilayer interoperable architectures, especially in climate and public health domains. During the "temporary suspension of regular data collection", the focused collaborative implementation with local partners of layered, explained, transparent and reusable digital twins of populations, people, and systems is an opportunity to be seized now and prepare for "better, safer, health systems, and planetary health in a post COVID-19 world". Examples and project themes may be broadly grouped into Health AI, and Climate AI.

The Health AI applications revolve around the Resident Safe Cities data assembly and the City Health Score, using local health data from the addressable resident population with local customized actionable intelligence and advice. Variable-source and variable-destination SCOR "Smart Cores On the Road" pace and modes composition data is collected and analyzed to inform advice about prevention and possible amelioration of the major burden of preventable disease and reduce the diseases of desolation such as

anxiety, depression and loneliness. Scalable, explainable, and understandable recommendations can be produced depending on the data availability and destination. Community Hotspots are “color coded” and further recommendations are given when cities face spikes in empty space with no residents.

10.1. Health AI Applications

Health and healthcare are domains where scalable models have traditionally played a minor role. Perhaps the greatest deployment of AI/ML techniques in public health is prediction of health resource needs such as hospital capacity, intensive care availability or health need due to disease spread from proxy modalities that change more frequently than disease outcomes such as hospital utilization in response to flu or RSV. Recently, several modeling teams have transitioned from prediction based on inverse modeling trained with frequent proxy information to prediction with direct models that require no training or rely on very sparse training due to the long intervals between previously available disease data instances. Disease prediction throughout an episode is useful for helping coping resource allocation and health management decisions including case isolation.

Expansion of deep learning into emission reduction from people and businesses makes health impact modeling of emissions-attributed diseases and other exposure-attributed diseases high-impact health application areas that are, however, not yet well-scaled applications. Exposure-response models are commonly fitted to country-level exposure data, so they need to be dis-aggregated to health prediction resolution. Over and under-aggregation can cause large uncertainty in the disaggregated estimates of health impacts. Population scaling relationships for disease prediction may also be nonlinear or may differ across regions. Resume utilization is distinctive by country. Therefore, it is important to set separate parameters by country. These distinctions motivate careful consideration of how controllable factors differ across regions so that they can be incorporated in causal and statistical health models.

10.2. Climate AI Applications

Climate AI systems are AI applications deployed to address challenges of climate change. They are also part of the general family of AI systems applied to climate science and geography, inexpensively automating many labor-intensive tasks within these fields.

Climate Change Monitoring from Space. A new generation of Earth observation satellites are providing the remote sensing data required to support Climate AI services at new levels of detail. Satellite systems are providing daily or frequent revisit coverage of Earth at high resolution. This data is being used to help governments,

insurers, non-profit groups, and scientists understand where present-day climate models are failing to properly incorporate human activity, and to help improve climate model projections into the future. Using these higher resolution observations will better calibrate models of land-use change, agriculture productivity, forest biomass, greenhouse gas emissions, and other systems highly dependent on local human activity and practices.

Climate-informed Decision Making and Reporting, at scale. Increasingly governments, corporations, and enterprises are under incentives to become better corporate citizens, to report on their own emissions, and to build out sophisticated emission-reduction plans. In turn, they have to process vast quantities of climate data, including satellite observations, weather and climate forecasts, climate projections, and more. Climate AI systems can help in structuring markets for carbon attributes, or in directly governing the reporting, verification, and reconciliation of carbon attributes stimulating the International Carbon Markets, estimated to become a trillion dollar market.

10.3. Lessons Learned from Implementations

There is no lack of scalable architectures or systems for AI hardware and software in the world today. In the area of AI operations for third party cloud services, there are companies that offer specific and unique solutions for model training, model serving, data storage, or data labeling; companies that specialize on the optimization of AI algorithms; and companies that provide solutions to speed up the deployment and operation of AI-enabled applications. Here we only discuss a few lessons learned from our own implementations of AI solutions for health and climate problems.

Multi-modal systems are easier to build than single modal ones, as more mature and more functional interfaces are available for end users. For several projects with air quality public dashboards, for example, we were able to obtain visitors feedback through image overlays and visual recognition, instead of explicitly asking for their opinion about the air quality on the area. Latent spaces learned from one modality can be a good initialization for other modalities. For example, early versions of a project were based on a computer vision system trained only with images of faces wearing or not wearing masks, and able to segment the whole image area with head detection and mask wearing information in the latent space; It was re-trained with user tissues and health behavior data, and then supplemented with real-time hospital admission and mortality information; And adapted for the several user groups and geographical areas, until able to provide accurate and reliable local predictions. Another example where latent representation knowledge transfer worked well was during implementation of deep learning tools for the matching of names in public records data sources. A template matching tool for image segmentation of handwritten documents did an

excellent job to suggest similar records generated with different forms or spelling variations; And even for different recorders handwriting.

11. Conclusion

AI has the potential to be an essential tool to help society deal with unprecedented challenges associated both with climate change and the many aspects of the related intricacies governing the health of the population. The scalable architectures, resource utilization and research agendas defined in this text should be taken into account when developing innovative algorithms, and applications leveraging existing and or novel AI methods to solve problems of great urgency and to create great societal and economic impact. The traces or footprints left behind by the efforts of developing such algorithms and applications as well as the datasets released into the public domain for the sake of speeding up present and future technological innovation must comply to both business ethics and regulation so that they accelerate progress and do not exacerbate the inequalities linked to the digital divide.

We also encourage researchers and companies from all AI-related sectors to take on board and act upon the advice provided regarding explainability and transparency of the intelligent assistants they conceive and deploy and the AI-supported decision-making flows into which they are integrated. The deployment of intelligent assistants in sensitive domains linked to our daily life that other industries have irresponsibly postponed and or resisted should be validated beforehand through a variety of processes that require the involvement of experts in the six stages mentioned: (Human) problem formulation; efficient and effective data collection; intelligent assistant train-test; validation and deployment; public and private exploitation; and monitoring and surreptitious improvement.

Above all, we invite the readers of this text to join forces and collaborate. Only together can we overcome the many barriers and obstacles currently preventing an effective use of AI for Health and Climate. Only together can we work upon what is currently a molecular approach rather than a systemic one in the attempt to leverage the many scientific and technological capacities Artificial Intelligence is empowering to tackle the Health and Climate challenge at scale.

References

[1] Ueda D, Walston SL, Fujita S, Fushimi Y, Tsuboyama T, Kamagata K, Yamada A, Yanagawa M, Ito R, Fujima N, Kawamura M. Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future. Diagnostic and interventional imaging. 2024 Nov 1;105(11):453-9.

[2] Kodakandla N. Scaling AI responsibly: Leveraging MLOps for sustainable machine learning deployments. International Journal of Science and Research Archive. 2024;13(1):3447-55.
[3] Sai S, Chamola V, Choo KK, Sikdar B, Rodrigues JJ. Confluence of blockchain and artificial intelligence technologies for secure and scalable healthcare solutions: A review. IEEE Internet of Things Journal. 2022 Dec 29;10(7):5873-97.
[4] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.
[5] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.
[6] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.
[7] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.
[8] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.
[9] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.
[10] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.
[11] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.
[12] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.
[13] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.
[14] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.
[15] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.
[16] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.

[17] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.

[18] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.

[19] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.

[20] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.

[21] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.

[22] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.

[23] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.

[24] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.

[25] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.

[26] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.

[27] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.

[28] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.

[29] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.

[30] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.

[31] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[32] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[33] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[34] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[35] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[36] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[37] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[38] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.
[39] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.
[40] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.
[41] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.
[42] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.
[43] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 11: Experimental Methodology and Validation Strategies

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction to Experimental Methodology

Experimental research is used to evaluate changes and their effects on a given response. In the expression change, it is meant the change of the values of one or more factors, whose levels can be modified, which is normally carried out in the process where the evolution of the response is evaluated under all those conditions where it should behave differently. The access to the conditions where the response behaves differently is in practical applications unfeasible, because it would imply the consideration of errors of considerable magnitude [1,2]. Nevertheless, it is possible to create controlled conditions in observations, for instance, modifying block parameters that are previously defined in their levels for a period long enough to evaluate changes in the response. Each change should be assessed with a minimum number of repetitions.

Experimental research can be performed in isolated experiments or a set of experiments that are jointly evaluated. The first ones evaluate problems at design time, but the second ones allow assessing instructions provided on line during the operation of the system. The isolated experimental methodology is more developed than the joint experiments one, that are usually performed by sets of replicated space-time observations that are used to eliminate the influence of randomness by their average over time. Both methodologies can be used to estimate models that are going to be applied in order to control the operation conditions of the process to keep its response at the levels that prevent undesired events caused by the errors in the influential factors or to minimize the effects of their changes.

Non-systematic experimental questioning procedures are often the only available way to improve the behavior of complex systems, since other available techniques are not easily applicable or simply unfeasible [3-5]. However, performing experimental research in the real world raises a fundamental issue. In existing systems, all factors of the systems are involved in the observed response. Actually, the response is influenced by changes in the levels of these factors. Assessing the joint effects on the system response of all these factors, some entirely foreign to the particular subject of study, would require an impractical effort to apply all possible ways.

2. Cross-Domain Benchmarking

Despite the broad applicability of deep learning in geoscience, enabling data-driven discoveries from sensor observations, core tenets of experimental methodology are still observed only in select studies – with the majority of models developed in domain-specific silos. In the scientific craft of developing a new model, common pitfalls include overfitting, lack of reproducibility, claiming superiority of models that share the training dataset, and missing truly independent test datasets. Cross-domain community efforts aim to remedy these gaps by providing platform-agnostic downloadable model weights; also including peer-reviewed experimental methodology and validation strategies. Going further, cross-domain benchmarking studies are particularly valuable as they address the very real concerns of stakeholder communities that already exist in silos, including finance, energy, transport, ecosystems, agriculture, and health – in doing so forming unexpected interdisciplinary collaborations among scientists. To illustrate these tenets of software and experimental methodology and validation strategy reuse, as well as scientific collaboration, we detail two cross-domain examples, and benchmark them against five geoscience-specific studies that complete the broader paper overview.

As first-hand examples of cross-domain benchmarking, recent work translates state-of-the-art deep learning models from the computational efficiency and accuracy of natural images to the rendered images of 3D tomographic reconstructions of patients' lungs, labeled in accordance to the COPD diagnosis by medical experts, and downscaled through nearest-neighbor interpolation. Other researchers from disparate institutions develop a convolutional neural radiative transfer-specific architecture that outputs the accuracy of atmospheric surface solar radiation products collocated with surface measures as a function of visible downwelling solar flux for 170 locations over Mexico across five years. Because each chosen location-to-collocated satellite pixel association at any datetime provides the radiative transfer input, in doing so, they immensely train the deep learning model. Performance fingerprints are generated, which quantify the varying influence of model inputs to the differences in accuracy across locations and dates.

2.1. Medical Datasets

Many medical image analysis methods have been developed in the last two decades. A majority of them provide either quantitative or qualitative comparisons within a single dataset domain, such as MRI brain images of patients with a certain disease. Binary segmentation is the more concentrated on task, as a result achieving higher accuracy. This conduct has provoked an exploration concept bias effect that results in overfitted algorithms.

To overcome this domain bias, certain authors have proposed to expose algorithms to multiple datasets in the training stage. The key idea is to utilize data from diseased patients of diverse origins, scanners and sequences. Unfortunately, few are the works that provide benchmarking protocols on the latent space. The reason behind is that pre-trained networks tend to implode, limiting the exploration of the latent space. We benefit from the Boosted Dataset Curation (BDC) digital image processing task that was proposed to reduce data heterogeneity. This task consists of generating synthetic data using Style Transfer (ST) neural networks that are pre-trained on other data domains, e.g. the natural image data domain. This paper goal is to investigate the ConvNeXt medical image segmentation task.

The major shortcoming of the state of the art studies is the fact that they only utilize internal datasets without conducting domain adaptation methods nor exploring cross-domain pretrained encoders. Furthermore, the proposed Boosted Dataset Curation (BDC) benchmarking lacks the search for the best resources. In this work, we investigate the internal dataset size; the encoder type and the input overlap ratio (IOR) for the ConvNeXt based medical image segmentation task for two different architectures with, respectively 25M and 88M parameters. We focus exclusively on the Contrastive Learning pre-training phase as it is the more tutor-inductive phase (i.e. it provides better inductive biases). The scientific features of the ConvNeXt architecture are related to a series of architectural modifications proposed in Vision Transformers for better tailoring images. These changes are replaced from the Conv2D-Maxpool block to Boosted CNNs (BCNNs).

2.2. Atmospheric Datasets

The atmospheric domain consists of different subdomains like thermodynamics, radiation or cloud physics. Here, we demonstrate our transfer function on selected atmospheric variables simulated or observed worldwide over the last decades. The data can be used for validation of new modeling approaches or perform cross-domain assessments. In state-of-the-art numeric weather or atmospheric reanalyses models, different physical processes are solved. Due to the different physical solving

approaches, correlations between predicted and observed data are on a lower level for atmospheric models than expected. With the presented transfer function, narrow mapping functions for model parameters can be derived, so that a post-processing step could bridge the gap between model and observed data, reducing further discrepancies. Forecasting or analyzing weather data is crucial for our everyday life, starting from when to take an umbrella with us, to how much snow to expect when going skiing. Physically based models for atmospheric simulations have a strong tradition supporting foreign applications like machine learning or deep learning. Still, the output of atmospheric physics-based models can lack realistic detail compared to high-resolution observations. The purpose of the Transfer Function is to estimate expected high-resolution observations from the output of a low-res and/or low detail physically-based model for atmospheric simulations, supporting various applications, like the reduction of perceived clouds from low-res satellite data. For the probabilistic Transfer Function, we show results for 6, 12, and 24-h lead time for precipitation, 2-m temperature, 10-m wind, and TOA radiation (upward and downward).

3. Evaluation Metrics

In supervised classification problems, the main goal of performance evaluation metrics is to allow the analysis of model prediction accuracies and thereby provide guidance to the user on the selection of models based on their specific needs and requirements. Performance metrics are usually related to types of costs associated with misclassifications such as computational expenses or misclassification penalties. The general measure for classifier accuracy is the proportion of verification samples that are correctly classified. The following equations provide a set of accuracy metrics for evaluating classification performance and are widely used for classification studies. Let a set of S denote the samples used for performance evaluation, in which N is the number of samples in the set S. Also let a set of P be the set of corresponding predicted labels of the samples in S. The accuracy, ACC, sensitivity (or true positive rate), TPR, and false negative rate, FNR, can be computed as follows:

where GT is the set of corresponding ground truth labels of the samples in S, and Ctotal is the total number of unique labels in both the predicted and ground truth sets. The accuracy metric can be used to provide information on the ability of the model to generalize over the training data (i.e., no or low overfitting) as well as the hold-out test set if a suitable number of tests have been performed. It is advisable not to base the performance of a model on a single run of the model on the hold-out test set. The true positive rate metric can provide insight on the ability of the classifier to correctly predict the presence of different classes. It provides information to the user on possible misclassification events that may exist for different classes. Possible misclassification events can be analyzed from the ground truth labels by delineating classes based on

coordinate conditions, or other pre-identified conditions associated with the input features.

3.1. Accuracy

Many system tasks are evaluated on how accurately they reproduce their intended output. The simplest form of incorporation of systems evaluation through accuracy is through the correctness of the perceivable output. For example, in generating visual images or 3D shapes, the accuracy can be evaluated by directly measuring the perceptual distance with that from the ground truth. Most algorithms share a similarity in this respect and many ground truth-based datasets exist. The distance metric used for evaluation is typically an average of representative samples, across a closely sampled grid or random sampling, with the average typically summarizing the geometric Euclidean distance of data points from the ground truth or the perceptual distance such as the Structural Similarity Index, Peak Signal to Noise Ratio for visual tasks or Point to Plane / Point to Mesh distance, Mean Perceptual Distance, Chamfer Distance for shape tasks. Some functions, that depend on the type of discrete correspondence used, can interpolate between the two. One metric is often defaulted without consideration of the task.

Nevertheless, current methods are unable to approach human capability in many modes of generative tasks across a variety of generic and category specific datasets. The growing disparity between human capability and automated system accuracy creates a need for a specialized system to avoid failure mode collapse since tendency for bias sample is a common occurrence in earlier attempts of datasets where learning capacity was limited. Often a similar sample shared with the ground truth becomes the default with a large disparity between the rest of the samples in the dataset and in high accuracy mode of such methods, convinces the evaluator to believe the generated sample to be of high quality. All these factors need to be taken into account and while accuracy remains ground truth dependent, the task remains the deciding factor for importance of accuracy in evaluation too.

3.2. Area Under the Curve (AUC)

3.2. Area Under the Curve (AUC)

In general, supervised learning models learn a mapping from labels and features. The problem is that the number of different possible labels can be really large; for example, in a typical talk, the number of different ways a speaker can say one sentence can reach numbers on the order of hundreds of thousands, simply because there are so many possible co-articulated phone variations in any language. So far in supervised learning for speech recognition, it is not possible to evaluate the quality of the scores of the

learned models because we need a particular correct label for each input. However, there exist more metrics to evaluate models than just the WER; in our case, to enable also the early selection of promising models during training, we use the number of best hypothesis model counts and an incremental extension of this metric, called the AUC. The AUC has already shown to be one of the best supervised learning feedback metrics in the case of unsupervised deep learning.

The basic assumption is that to predict the output class of input data with a supervised model, sequence-dependent feature maps are learned, mapping speech data into the space of possible labels for a specific word. The different phases while computing the AUC consist of iteratively computing audio features with the lowest data representation error. When the optimal parameters of the speech data are found, the model is used to check the common phonetic coincidences of the different possible predictions of the model. The output is the word with the highest counts in the space for each input audio feature, thus called the AUC (Area Under Curve) metric.

3.3. Interpretability

Interpretability is one of the most important properties of computational models, and is usually defined as the degree to which a human can understand the cause of a decision. Models with increased interpretability can improve trust in their usage, while less interpretable models may require more extensive validation to address risk mitigation across key stakeholders. Indeed, in models requiring a high degree of trust, decisions should be testable and explainable by model creators [6,7]. Work has demonstrated that users are typically satisfied with explanations provided by especially interpretable models, and that they lead to greater trust in the model's behavior. Naturally, the most interpretable type of model is one that is fully understandable to a lay audience. Conversely, complex or black-box models are typically considered not interpretable. According to these definitions, our intent is to select an experiment approach capable of producing an interpretable predictive model. Experimental approaches that produce surrogate predictive models are generally more interpretable.

Interpretable models are also required to produce interpretable results, including coefficients and error metrics, testable hypotheses, and testable model designer assumptions. The concept of verifiability is synonymous to interpretability. A more complex model is more verifiable if it replicates the results of a simpler or lay model. The concept of local explanation translates to interpreting groups of training data. Explanatory tools exist that construct visual representations of the decision regions of certain classifiers, enabling interpretation of binary classifications across feature values. For example, counterfactuals-users-explanations explain a model's decision by

describing a user, or stakeholder, group as in need of special interventions to assist in preserving their employment.

3.4. Robustness

It is desirable that classifiers are robust not only when the feature set is modified, but also when certain properties of the training set, model, or classes in the model are modified. For example, robustness of a model is usually evaluated over a set of samples not included in the training set and presenting a different distribution. For a model to be robust to a modification of the sample distribution means that the performance does not significantly drop, and is measured using the performance metrics of the classifier. Usually, the validation set is the one used to control for these differences, but such modifications may create problems if the validation and test sets are separate. Inverse Probability Weighting has been introduced to manage different distributions over the training, validation, and test sets.

Due to the complexity of real world data distributions, as well as the computational and human efforts, the size of the training set is much smaller than in other supervised learning applications, such as image classification [2,8-10]. Therefore, several techniques have been proposed to increase model robustness for classification of small samples. Given the size of test sets used for classification, it would be unreasonable to propose that robustness of a model to perturbations of the test set should also be demonstrated. Nonetheless, it is simple to perturb the test set for certain problems, such as in cosmology, where image quality is controlled by the Point Spread Function and blurring, contrast, and noise perturbations can be easily made.

4. Clinical Trials

When a medical device has successfully undergone experimental methods but has not been manufactured and tested in the field, it has the status of an experimental medical device. Experimental medical devices retain this investigational status until they have successfully done so in the field during clinical trials. Clinical trials differ from other forms of research in that they typically take place after a medical device has reached the level of maturity specified and demonstrated effectiveness at the accuracy and/or precision level specified in the Standards against which the device will be evaluated by the appropriate regulatory body [1,11-12]. They differ from the research described in prior sections in their methods, purpose, and outcomes of study.

Clinical trials assess whether a device can perform well when used by the intended population as designed, and not just when used by someone who is well-trained, using specialized protocols in a controlled environment. Tests are designed to be representative of the actual intended user and clinical environment in which the device

will be used, with all of its benefits and limitations of intended use. Because of these differences, clinical trials differ from other research studies in their design, protocol, execution, and data analysis. Standardization of procedures is difficult, and factors affecting the outcome may be difficult to control. Therefore, stated goals of the study may go beyond efficacy and efficiency, and include sensitivity and specificity, ease of use, user interface acceptance, or level of work. Despite these limitations, a clinical trial is the most effective and efficient method of determining whether a medical device is ready to be marketed.

5. A/B Testing

Perhaps the oldest and simplest experimental methodology is A/B testing. In its simplest form, A/B testing consists of assigning test subjects into two groups. The first group (the control group) is shown the current, baseline example of whatever is being studied, while the second group (the treatment group) is shown an experimental example that differs from the baseline in a controlled manner. After the test subjects interact with the example they were shown, a metric related to performance is computed and used to determine how the two groups of test subjects compared, and thus whether it is likely that there is a difference between the two conditions.

A/B testing is used to measure differences in a performance metric between two versions of a product or service. For instance, the conversion rate of a webpage could be the metric of interest. Problems are typically small in scale and short in duration. Because running A/B tests causes test subjects to be assigned to one or two conditions with a random assignment, confidence can be easily calibrated using statistical sampling methods. When done routinely to evaluate changes to a product or service, A/B testing is a form of experimentation that delivers intermediate feedback during the development process and thus enables rapid learning.

6. Stakeholder Validation

Some validation processes examine the artifacts generated in the activity of designing artificial systems. Other validation processes interrogate the adequacy of a particular methodology with regards to a particular domain. Both sorts of validation are ultimately shallow, epistemologically speaking [13-15]. They do address the questions of whether the methodology allows for the successful generation of artefacts, and whether it is useful for some applications. The first question is an empirical one concerning the relationships between the methodology and the activities of the artefact designers. The second states a formal requirement for methodological soundness but cannot be addressed by methodology design alone.

While addressing these two validation questions is important, they are fairly routine. They put the focus on the methodology as a tool for artificial design. Concurrently escalating in difficulty is the question of whether the scaffolding provided by the methodology leads to any actual improvements in the activity of designing artificial systems, and, especially, whether these potential improvements constitute the sort of increases in wealth, welfare, and satisfaction of stakeholders that sociopolitical and ethical tenets prescribe for the systems developed through this exercise.

Difficult as it is to assess the impact of a particular methodological framework on the work of design, some meta-methodological contract designs demand that the methodologies they elaborate be designed for the stakeholders, in a responsible and ethical way – the purposeful design of the social scaffolding into which the developed systems are ploughed, upon which they rely, and with respect to which they should be validated [16,17]. In this sense, the responsible scaffolding of socio-artificial systems lies beyond the mere adequate functioning of the artificial systems themselves.

7. Comparison of Methodologies

In the previous sections, we have described selected methodologies for exploratory validation applied to potentially latent and predictive models. The selection reflects our subjective opinion and is not limited to the reviewed methods. In this section, we describe the main methodological differences and similarities.

In the comparison of methodologies, we include a summary of criteria that potentially define their applicability, advantages, and disadvantages; validation-oriented variants of modeling tasks; and an underlying model framework. The summary of the characteristics pertinent to exploratory validation is presented. The validated modeling task correspondence is highlighted [12,18-20]. The comparison of the considered methods is portrayed.

Now we briefly discuss mapping the examined methodologies to the model validation tasks summarized and elaborated. All validation methods can be considered and applied as extensions of model building methods. Statistical data mining methods, as well as visualization approaches based on clustering, outlier detection, linear regression, and local modeling tasks can be utilized as a potential alternative and supplementary answer to the predictive modeling task. Certain algorithms can be considered possible answers to the class-imbalance handling problem in predictive modeling, albeit with certain limitations. The models of local modeling, clustering data, and some outlier detection methods could be used as an answer to the explanation of concept distribution problem. The model agnostic validation approaches are representatives of information extraction tasks.

8. Challenges in Validation

There are further challenges to validation that are tied to the way the data collection and analysis processes were originally set up. The first is that in the family of work that usually falls under the label of qualitative research there is an explanation of how the data was collected and analyzed as part of the study but often the 'validation' processes are largely absent [21-23]. For example, take the classic AAR. It is a written record of the details of a group discussion that was facilitated to an uncontrolled or too poorly controlled degree. That record is of group opinion at that particular point in time so not only is it readily interpreted in numerous ways but the ideas raised are susceptible to distortion and even distortion by the researcher in his or her subsequent analysis. Further confirmation is hardly ever provided. Another simple idea in qualitative research is triangulation—that is, use of multiple sources, methods, investigators or theoretical perspectives—should increase the accuracy of the findings. Such triangulation is rarely found in qualitative studies. Great promises of insight into our concerns by qualitative methods are often not reciprocated by concern for validation.

A related challenge for these qualitative methods is that whilst they have become the go-to-place for many ideas of practice designers and practitioner instructions, qualitative interview programs for the objective of insight generation are remarkably absent from the qualitative literature in the main [24,25]. This may indeed be a reflection of the starting point for many of the methods—user testing, redundancy testing, instructions, prompting, feedback—that are primarily explicative in the objective they serve. Indeed famous qualitative researchers writing on practice explain these methods solely in terms of the explicative and then propose a conversion process design that uses interviews from a validation perspective. In contrast, however, user tests, focus groups and the like, are widely used in practice for the validational objective [26-28]. Yet the qualitative literature has created a kind of market within practice—designing for qualitative work and validating, often at an entirely different level, with quantitative methods. Enhanced by the use of qualitative methods described by many researchers as being exploratory on an entirely different level of empirical concern.

9. Future Directions in Experimental Methodology

This chapter emphasized a general framework of strategies that can be employed for carrying out experimental research in psychology, psychiatry, and neuroscience. The goal of introducing this Experimental Methodology and Validation Strategies framework was to provide guidance for the implementation and assessment of experimental research strategies and to help improve their quality and previous

validation. Yet, as is often the case in science, although introducing a methodological framework is a first step towards improving the methodology in future research, questions continue to remain.

In the introduction to this book, we quoted a historical figure who claimed that "nature is a book written in mathematical language." Our view is that to make this language understandable, we need the right experimental tools and approaches. In this final chapter, we will devote some time to analyzing possible future developments in experimental methodology and suggest possible future experimental tools that have not yet been developed for psychological, psychiatric, and neuroscientific basic and applied research. We will address three different questions. Firstly, what new experimental tools and approaches can psychology and psychiatry borrow from neuroscience? Secondly, what new experimental tools and approaches can neuroscience adopt from psychology and psychiatry? Thirdly, how can all three disciplines bridge the gap between applied and basic research by devising experimental tools with possible translational value?

10. Case Studies

A fundamental goal of science is to arrive at universal, predictive principles. However, advances in knowledge typically begin with the careful watching, measuring, and recording of some noted deterministic interesting events and relationships, i.e., phenomenon. The experimental researcher takes the first steps toward advancement of knowledge by attempting to predict or explain or better yet, elegantly simplify the impressive complexity of natural and societal systems. This is a significant activity in science. Case studies serve several important functions. After the experimentation and hypothesis building stage of research, case studies precede the quantification needed for theory formulation. As a scientific process, research is fundamentally iterative; the theory, experiment, and case study phases are mutually informative, feeding off one another, leading to improved calculations and models. Thus, case studies often precede experiments. This is clearly seen in engineering, where prototype construction and testing serve to refine concepts and formulas. Of course, the case studies may also be done in a later phase, serving the purpose of model validation, where the design and implementation efforts are greatly aided by documenting and understanding existing systems. In applied dynamics, the predictions of models based on theoretical core concepts are confirmed or refined by comparison to the measured behavior of a repository of “real-life” systems.

11. Ethical Considerations

Research in the area of human-computer interaction is an experimental science, to validate or develop fundamental principles requires understanding the whole system

and often much experimentation [29-31]. There are therefore ethical considerations associated with any research that relies upon interactions with people, who may have to provide responses under time pressure in unnatural conditions. The expected benefits may be covert, where the participants are unaware that they are participating in a test, while at face value the participants may not realize that they are engaged in experimental work. Recent changes to societal views about such techniques suggest that further discussions may be entered into regarding tolerability of covert task-based experimental work. In particular, there are groups in society that are more sensitive to exploitation. Vulnerable groups include refugees, children, elderly citizens, individuals who have mental disabilities or who are suffering from other shocks. For the reasons noted above such research demands the highest ethical considerations on experimental design, and to-date there are very few guidelines available [3,32,33]. Current guidelines focus on organization at the institutional level and on human rights within the system.

As more technologies, from surveillance to recommendation systems, delve deeper into the social and behavioral fabric of societies, those interested in human-computer interaction must think critically about the implications of their work. From before the inception of a project until the end of its lifespan – including failure – interaction design is filled with the potential for harm, but often devoid of the ethical scaffolding to help guide researchers through those waters. Researchers are often unaware of the negative impacts of their work in the short, medium, and long-term and may not prepare for these eventualities. Interdisciplinary work with areas such as design ethics, ethics of care, critical design practices, provocative design, or concept-driven design can help address ethical questions.

12. Data Collection Techniques

A collection of various techniques, as described below, was employed for information gathering. A particular department in a private business school in central India which conducts training for various professional and job oriented courses was chosen as the case under study [4,34-36]. This institute also operates one of the B-schools approved by the Government of India. However since it is relatively small operation with a limited scope, the research questions are best answered through a busy service perspective or observations by relevant members, thereby helping meet the research objective.

Discussions with the director of the institute were first held to establish an understanding of the industry scenario and specific challenges in managing such a training environment. Executive coaching was identified and defined with the help of this discussion and results of the preliminary interviews. A detailed interview guide

was developed, drawing on relevant academic literature and primary interviews conducted previously to highlight important issues to be discussed. Pre motivation training perspectives were next obtained from sixteen members involved in career and leadership development processes of the coaching [37-40]. Responses to this guided set of face-to-face interviews were recorded, transcribed, and coded using software to identify specific patterns and themes to design a more general study from the patterns in the data.

Interview results were first reported separately for each specific respondent category. Furthermore, implications from the qualitative expert-exchanged knowledge were also detailed and followed up through an informal discussion with selected members. This provided further realization of the perspectives of the mentors on specific implications and parameters of predictiques.

13. Statistical Analysis Methods

Statistical analyses of bioscience experimental data are common. Most experimental results, especially in biology, contain noise, and statistical methods help address that. Such analyses help in understanding if an experimental result was just an outcome of random noise, or if there is a belief level associated with the result. Statistical evaluation using various tests is a standard part of experimental results today, as such analyses add confidence and believability to the experimental results [4,41,42-43]. For determining the impact of an experiment, data comparison is done primarily in three common ways: during the course of an experiment, or in assessing the final data outcome. Statistical tests are primarily used for the second methodology. Aspects such as sample size, normalization methods, repeatedly performing the experimental procedure for every condition, need to be factored in. Common statistics software packages include common statistical tests widely used are ANOVA, t-test, Chi-square test, and K-means clustering. Popular methods used for clustering experimental data together are the heatmap method and the principal component analysis method. These common tests are highly useful. Even so, it is important to realize that these tests are meant to address a single aspect of the variability present in the experimental data, and other methods and robust confirmatory tests should be done. Often there is no clear and simple method available to choose from, or the correct statistics cannot be done, to analyze every type of experimental data outcome. Therefore, it is essential that careful testing and choosing analysis methods that make sense for all aspects of the experimental data are chosen.

14. Software Tools for Validation

This chapter discusses computer programs that help people conduct some parts of validation. Some software tools integrate many validation activities from all stages of

the research process. Acknowledging that validation has a high degree of subjectivity, others are built to deal with particular aspects of validation, supporting or automating the decision-making, which are generally performed by the researchers. In both situations, the collected results, which help or automate the researchers' decisions, also need verification.

To the best of our knowledge, no other software manually conducts, supports, or automates the type of empirical validations previously mentioned in this chapter. Available software tools provide only basic mechanisms to inspect artifacts that harvest the kind of results expected from validations shown in this chapter. Such tools' existent capabilities partially automate traditionally manual support activities. Thus, we detail options for developers interested in building validator assistants. For some activities in this chapter's descriptions, specialized software exists; some are macro programs embedded in commercial tools developed for other purposes, which are then adapted for validation.

Additional software is available for the attributions of information extractors, comparing results with multiple extraction methods, and statistical analyses. Many of these validators work with specific model types, such as defect prediction, mutation test, and sentiment analysis models. These tools incorporate the most used validation methods on prediction, testing, and exploratory validation methods. Nevertheless, proper use of these tools requires knowledge of the earlier detailed validation methods. Little or no user interaction is required, whichever the validation or the tool. These tools' results, such as the parameters for using models, need validation for information obtained from other results using validation methods.

15. Interdisciplinary Approaches

Although interdisciplinary collaboration presents its own challenges, it opens the door for potentially enlightening research and findings that can only come from a pool of experts in multiple subject areas. By taking on similar questions or investigating ideas that can only be understood from the results of other disciplines, researchers can realize the full impact of their research topics; for cognitive systems and sciences, working with those outside of related fields enables larger implications than previously imagined. Combining topic elements across disciplines is a natural and often foundational means of investigation and experimentation, driving more complex models and system functionalities than available given the limits of a single field focus. The implications of reasoning, cognition, and study are across all aspects of the human experience. The further researchers step into advanced methods and deeper understandings, the larger the impact remains. These pursuits should be unavoidable and encouraged, given that they broaden the implication pool for all.

The field of human-like cognitive capabilities development is still in an infancy state, only recently able to begin tackling problems that advance the range and depth of intelligence feats possible. The convergence of humans and machines at the level of complex behavior and action creates a view that requires justification for existence from both embodied in their respective environments. In looking at the areas surrounding both entities, it becomes possible, and probably necessary, to look at the topics in a focused perspective of both the human and the tool, considering them fully functional entities and systems. These ideas emerge from the evolvement and discussion of the relationship between humans and technology and how the joint actions within the environment define and understand both from a full aspect view.

16. Best Practices in Experimental Design

What makes a good experimental design? First and foremost, a good experimental design distinguishes between a controlled manipulation of the factors of interest, and the measurement of the outcomes. Policy evaluation is often about understanding the causal impacts of some action or event on other variables. These variables may often be of interest in their own right, or they may act as intermediaries to further impacts. Treatment and impact – two sides of the experimental coin – need to be defined. In the case of laboratory or field experiments, treatment is usually some sort of controlled intervention manipulated in a random manner by the experimenter. In the case of natural experiments, this usually consists of some specific set of circumstances that arise, but are not manipulated by the experimenter.

Impact consists of a set of variables which are not manipulated by the experimenter but are measured as part of the experimental process. In an idealized context, that is, we would want to make sure that the only differences in the outcomes across participants in an experimental setting were the result of the inputs manipulated within the experiment. This is not a trivial exercise. In particular, we would want to make sure that the treatment is not confounded with other factors that vary across participants. And this is why we conduct controlled manipulations in a laboratory or other controlled settings, or use randomization in field experiments. The ideal of a well-executed experimental design is one in which comparisons can be made without concern that results could be affected by other processes occurring outside of the experiment. In short, a well-designed experiment reduces the possibility that the observed results are affected by confounding factors.

17. Limitations of Current Strategies

Despite the wide use of experimental research methods and their increasing popularity in various research fields, there still seem to be several limitations concerning their application. In this section, we will elaborate on some of these limitations. It seems

that most of the current experimental work has focused on its scientific rigor at the expense of its naturalness and potential contribution. Drawing on issues raised in interpretive research, we argue that experimental researchers should take the context of the study, the emotional reaction of the participants and the short duration of the Q&A into consideration, in order to enhance the naturalness of the study and to contribute to additional outcomes, additionally to cognitively evaluative outcomes. Moreover, we suggest that research contributing to a Q&A pattern would greatly benefit from careful longitudinal field work in real context, informed by insights from experimental methodology and its focus on delineating connections among phenomena of interest. In our review of previously conducted work in experimental methodology, we note that there has been greater emphasis on cognitive and less on emotional outcomes, perhaps because of emotional outcomes being arguably harder to study using experimental strategies. However, in the context of the traversal of agential transformations inherent to answering a question, other than merely event outline or closure types of outcomes, we suggest future researchers must explore the trajectory of emotional marking that makes for a complete response to a question addressing anticipated informational need. In short, we see an identified gap in contributing to emotional outcomes using experimental research methodology. We point out that some of the tasks inherent to question and answer exchanges, such as asking, directing, or justifying are moral, as they imply that one member of the portfolio distribution retains the right to dictate how an event will be played out. This introduces power relations to the research project, which due to the nature of this type of research, is very difficult if not impossible to eradicate or mitigate, as it is usually the researchers' presence that elicited the Q&A dynamics they measure. Finally, we note that while experimental approaches are predisposed towards uncovering cause-effect relationships, the type of claims researchers can make heavily depends on the context and the nature of the stimulus used, which is something researchers applying experimental strategies must always take into consideration.

18. Integration of Findings

One of the basic tenets of an experimental study is that key findings of the study must inform the real world. The classic way to do this is to generate post hoc conclusions to the effect that, for example, "this represents progress towards the study of some interesting aspect of the real world" or perhaps "these results may help inform decision making in some interesting application domain." Increasingly, however, experimental researchers are feeling the need to go beyond such post hoc claims. They are making intentional efforts to develop research agendas that derive technology, design, or other social implications from their research findings—that tie their findings into a larger whole. Of course, not all experimental studies have to attempt these ambitious goals; this level of integration is not a requirement for an experimental study. In fact there

may well be good reasons not to undertake such integration within a given experimental study, either because the study is one of a series of studies working towards some eventual goal, or because there are other mechanisms for integrating your finding with world use—for example, you may generate an external proposal mechanism, like a framework or theory, that others will use to make that tie-in. This section describes some of the techniques that experimental researchers are using to bootstrap their research and not simply rely on the good fortune of being cited.

19. Feedback Mechanisms

The common view is that scientists and engineers receive feedback on a project's conformance to requirements only at the end of a project, as manifested in product test and evaluation activities. The purpose of testing is to see how well a product meets the customer needs. Feedback on the engineering design is only one aspect of product testing. The other aspect is a check on the correctness of validation or experiment planning. Implementation tests should first demonstrate that a significant experiment is feasible in the method being considered. When that is accomplished, the final experiment can then be executed. Conformance tests confirm that the product meets requirements but have a special significance for the engineer. Those conformance tests can point the engineer in the direction of errors in the engineering design. They provide alternative explanations for the apparent problem during the engineering design process known as the "happy path".

Giving feedback to the scientist is much more complex than providing feedback to the engineer. The scientist must be allowed the creative freedom to explore; to risk making an error that will fail validation in order to advance knowledge. This freedom must be governed by an understanding that validation is a more sensitive tool than conformance testing: at least errors in the design of the validation event carry many more implications than an error in an aspect of the engineering design. It is our belief that the first step in structuring feedback on validation planning is to stress the importance of such planning and to develop early test plans with the scientist. The plan must be more than a simple listing of the aspects of validation and the design of studies to be employed to study them. That plan can easily be too long and intrusive to be practical. What we propose is an understanding, formal or informal, that is recursive but escapes being an unwieldy document.

20. Summary of Key Insights

This chapter concludes our thesis about the role of experimental studies in empirical software engineering (ESE). It urges researchers to deliver useful research-based guidance on how to make decisions. Big data analyses yield many guidelines addressing the question of what to do, but few addressing the questions of why or how

to do it. This chapter proposes a new type of empirical study, experimental confirmation of observational studies, whose purpose is to contribute to our understanding of ESE and how to deploy it more effectively. This chapter also presents eight guidelines to improve the practice of experimentation in ESE. The proposed research direction is valid independently of whether ESE is considered a real engineering discipline or not. The guidelines are equally valid for ESE whether one considers it to be in a jungle state or not. Experimental confirmation of the results of observational studies can benefit developers of both academic and industrial applications of ESE, irrespective of the particular applications being developed. The ideas and guidelines discussed in this chapter are informed by our combined experience of formulating, conducting, analyzing, reviewing, and discussing empirical studies in the area of ESE.

Although ESE researchers collectively have many years of experience developing empirical studies, we still make mistakes that lead to suggestions of questionable practical value. Many of the mistakes that we make relate to maximization purposes, study unit selection, lack of attention to noise, and moderator variables. These suggest several recommendations for improvement. To make a valid comparison, a study should include a sufficient sample size that justifies the allocation of units to different experimental conditions. The units should involve sufficiently homogeneous data points to avoid masking effects due to noise on achieving different outcomes. There should be enough noise control and/or correlation in the moderator variables of interest to permit their study if necessary. The conclusions should make clear the limitations on the conditions of practical value based on the study's design, duration, and data collection.

21. Conclusion

We worked through a variety of considerations and proposed some methods and techniques to those working in the Computational Intelligence field, first in experimental design and second in experimental validation. Everyone working in AI, Computational Intelligence, and any applied Databases needs to understand what is, and is not important given the state of our understanding. Such exploration should be done with little fanfare, and in such a way that it does not mislead those who will build upon it. That understanding is expressed in the details of design of an intelligent system, whether as an agent, a theoretical construct or something more abstract – differences that are often overlooked.

Too frequently the question of what exploration demonstrates, and the point it makes is never adequately explored. The gap between the new work and the earlier work should be underscored. A question posed in the early work should be asked. What was

experimental economy in getting to that point? Was it that the earlier work once made assumptions thought naïve but rather clever? Did it explore tiny corners to reach conclusions so general they need to be debated in fine detail? Why did no one earlier think of it? Design pride is why it doesn't happen more often. Comparison of performance amongst similar methods is error prone. Comparison of methods with distant similarities tends to suggest reasons on performance related to slight changes in the details. No one would propose leveraging an AI system to detect race conditions in programs, lift design from diverse branches, and insert it all together. No one would propose even lifting one class of function and then inserting them together again.

References

[1] Hartung T, Kleinstreuer N. Challenges and opportunities for validation of AI-based new approach methods. ALTEX-Alternatives to animal experimentation. 2025 Jan 14;42(1):3-21.

[2] Myllyaho L, Raatikainen M, Männistö T, Mikkonen T, Nurminen JK. Systematic literature review of validation methods for AI systems. arXiv preprint arXiv:2107.12190. 2021 Jul 26.

[3] Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, Lyngdoh NM, Das D, Bidarolli M, Sony HT. Artificial intelligence and machine learning technology driven modern drug discovery and development. International Journal of Molecular Sciences. 2023 Jan 19;24(3):2026.

[4] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.

[5] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.

[6] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.

[7] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.

[8] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.

[9] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.

[10] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.
[11] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.
[12] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.
[13] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.
[14] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.
[15] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.
[16] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.
[17] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.
[18] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.
[19] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.
[20] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.
[21] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.
[22] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[23] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[24] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.

[25] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.

[26] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.

[27] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.

[28] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.

[29] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.

[30] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.

[31] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.

[32] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.

[33] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.

[34] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.

[35] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.

[36] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.

[37] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.

[38] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.

[39] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.

[40] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.
[41] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.
[42] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.
[43] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

Chapter 12: Translating Artificial Intelligence Research into Impactful Solutions

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction

Artificial Intelligence (AI) has progressed in leaps and bounds over the past few decades, more so over the last few years, becoming an almighty tool in the arsenal of researchers and practitioners. However, translating these developments in AI research for use in society is a herculean task. It requires collaboration and compromise from all parties involved, researchers and solution designers on one side and solution end-users on the other. This 'technology push' and 'demand pull' needs to happen in the same direction to realize the potential impact of AI. Therefore, an essential function of the research community is also to provide technological solutions for real-world problems and to develop collaborations with possible end-users to deploy the newly developed solutions, primarily focusing on resource-constrained settings [1,2]. In the past, a few researchers have taken mainstream research trends such as AI for Climate Change, AI for Road Safety, AI for Disaster Recovery, AI for Public Health, AI for Mental Health, and attempted to align them with the efforts of the UN for developing a socio-economically balanced and environmentally sustainable world community.

At the same time, there is also a need for the collaborative efforts from domain experts from various fields, such as humanitarian aid, public health, disaster recovery, and others who develop the ground-level need-based solutions, and researchers from AI who collaborate with these experts and help with the solution design aspects [3-5]. The objective of this write-up would be to present the usual process for building and deploying a solution powered by AI, describe some of the use cases where AI was successfully translated into practice with examples, present the working of both collaborative and multidisciplinary aspects of this process through a couple of photo

books, and finally, discuss the present-day and future challenges in this endeavor, particularly from inter-disciplinary collaboration and diversity perspectives.

2. Cross-disciplinary Collaboration

A major theme in work to bring research in AI to bear on real world problems has been the importance of cross-disciplinary collaboration, synthesis, and translation. A key motivation for funders and impact-driven researchers is the desire to ensure that findings in AI lead to concrete and tangible impacts on challenges that humanity is collectively facing [2,6]. Specific areas have been prioritized as especially deserving of support, including climate and Earth Sciences, breakthrough discoveries in the biological and medical sciences, and newly enabled computational capabilities that mix and match both existing approaches and novel formulations.

Integration of Medicine and AI

Advances in quantitative and dense biomedical data acquisition are converging with rapid progress and readiness of machine learning methods. The challenge is to integrate the many modalities of bioimaging and other measurements enabled by decade-long investments in new-generation biotechnologies with epi- and meso-scale modeling of the complexities and emergent properties of living systems [7-9]. There are examples of clinical translations and realizable technologies but limited sampling and access is a recurring barrier in all areas including cancer, neurodevelopmental disorders, and neurodegenerative diseases. Possible applications in biomedical imaging include multi-modal and/or temporal domain synthesis, denoising, and segmentation as well as auto-regressive applied diffusion modeling.

Applications in Earth Science

The Earth Sciences are facing enormous challenges to understand a rapidly changing planet and its feedback on the living systems it sustains. Realizing the full potential of automated scientific discovery processes will require radical rethinking of both the input-output and closure relationships that connect observations to parameterizations/models that govern latent dynamics [10,11]. Physics-informed machine learning techniques might integrate new domain knowledge, efficiently harness diverse streams of geolocated image and time-series data, and interactively connect modeling and discovery to allow investigative discovery-driven exploration and learning.

Computing Innovations in AI

AI is driving innovation in the design of computing systems, component architectures, and stacks. Large Sensor Arrays enable unprecedented new confocal fluorescence

imaging of cellular tissue structures. New Algorithms and Hardware Architecture co-designs aim to capture up to a factor of a thousand improvement using Optical Computing principles. Many emerging applications in communication, computing, signal processing, and AI would may be better served via novel periodic nanostructured media platforms compared to current electronic and photonic devices, components, and integrated circuits.

2.1. Integration of Medicine and AI

Perhaps the most straightforward application of AI technology is to medicine. Most applications of AI in medicine seek to employ the ability of supervised learning to find mappings from some input features to a predefined output label to augment the ability of physicians to provide treatment [12-14]. These input and output variables vary by application. For example, disease diagnosis tasks seek to predict a disease label from clinical input variables. Disease stratification tasks seek to predict an outcome of interest such as survival time, disease recurrence, or treatment resistance at a later time. Treatment recommendation tasks seek to predict the ideal drug or therapy for an individual patient given a set of clinical features. Phenotype discovery tasks seek to attribute certain clinical features to a certain disease label for a group of patients.

The requirements for all of the labeled data used in these supervised learning tasks have arisen from real-world medical problems. Clinical data from millions of patients have been utilized to train, tune, and evaluate the predictive performance of the models. These models translate pilot studies in clinical practice and medical research into quantifiable statistics in clinical prediction, acting as important tools for decision making in clinical practice [3,15-17]. AI models relating to image or genetic data are especially influxing into medicine, as they can include input data that are challenging and expensive to collect, but greatly benefit from approaches. By combining the strengths of physicians and AI, we hope to decrease the time spent diagnosing infectious diseases or the path towards developing a rare disease and personalize the therapeutic strategy.

2.2. Applications in Earth Science

The quest to realize AI's potential for impact is nourished by a variety of interdisciplinary joint projects, such as climate modeling, weather forecasting, geotechnical applications, and resource monitoring for oil and gas. Whether through in-house innovation labs at traditional industry players or by startups created and performed by investors, the domestic resources sector is increasingly recognizing the benefits of adopting AI techniques to enhance and expedite resource identification and monitoring applied to geophysical and other properties characterization or prediction.

Peer pressure from successful industry or academic partners in geosciences will benefit resulting products quality [18-20].

The geophysical modeling capability within the SimNet tool combined with reliable AI technologies such as the Deep-Equations technology overlay for solving ordinary, parabolic, hyperbolic, and elliptic modeling differential equations open the way for other very different but equally useful internal R&D Joint Projects such as: improving the geophysical inversion speed and accuracy; provide a much richer set - multiple scatterers, compensated data - than that currently available, and use easily accessible geophysical observational data, very much like already being done with the electricity consumption data, to spur innovation and advances in the geophysical scientific adventure. AI techniques, including the DeepEq overlay in particular, realize unique capabilities and enable disruptive projects in Physics- and Mathematics-Informed-AI over the Breaking the Curse of Dimensionality, wave and other related industry standards; is the industry benchmark for complex 2D and 3D geophysical earth inversion and modeling. The Disruptive Physics-Informed-AI Breakthrough is obtaining such inversion results in seconds on a laptop risk-free.

2.3. Computing Innovations in AI

Advancements in AI have their roots in accelerated innovation in computer architecture, distributed computing systems, massively parallel architectures, and novel computing primitives. The co-design of algorithms and hardware has proven highly effective. Algorithm-friendly hardware architectures specialize in accelerating fundamental tensor operations [21-23]. Specific neural network layers may be optimized in speed, precision, or power consumption. Cloud-scale systems built on these accelerators, with high bandwidth between chips, nodes, and racks, now provide tremendous raw compute power in addition to model parallelism, data parallelism, and pipelined parallelism. These platforms have allowed researchers to leap beyond what might be possible with inference on surfaces and solved relatively quickly in the back end of the design pipeline and instead combine generative neural networks with inverse rendering, differentiable rendering, and diffusion processes. The future of real-time graphics and fabrication of novel products lies in integrating symbolic AI with deep learning generative models inside these accelerated systems.

The co-design of hardware and software algorithms is also the key to embedding large learning models into resource-constrained devices, such as cars, phones, drones, and robots. Custom silicon, with many architectural innovations such as architectural sparsity, low bit width data types, and mixed precision training, is critical for both fast inference with great user experience and low carbon footprint due to user-scale ubiquitous deployment. These devices may also leverage novel optimizations, like

pruning, quantization, model compression, and students-teacher distillation to embed large deep learning models into mobile or edge devices. Because the data-rich applications on the edge may induce new data distributions and makeup new requirements, the AI model will also need to self-adapt using low-volume streaming data.

3. From Prototypes to Scalable Platforms

The ultimate test of any engineering achievement is whether it is scalable, meaning that it is actually applicable to a global problem and can be executed in a cost-efficient way. Any engineering achievement could be conceived in many different forms ranging from something specific and focused on one instance of the problem, something generic and capable of applying to a myriad instances of the general problem – i.e. a business architecture – and something that arrives at the latter, powerful scalable solution incorporating the feedback obtained from a multitude of prototype deploys. These two approaches are aimed at enabling deployment at scale. A scalable solution is ultimately amendable to the architectures of businesses that aim to be the long-term major players. In a larger sense, companies and start-ups are responsible for translating the innovative research ideas into scalable market structures capable of servicing the demand at feasible delivery costs. In this sense the transition from research into companies that execute the ideas is the phase that pretty much dictates the impact of these novel technologies and has been sorely neglected.

The whole artificial intelligence research area evolved both around fundamental concepts of the nature of computation as well as around implementations of prototypes that perform impressive feats of human-like intelligence in a limited form [9,24,25]. The limited nature as well as the impressiveness of the prototype phase has possibly led some researchers in the field to think that a possible extension of the prototype behavior may allow at least some degree of scalable capability. Research funding may be directed towards one of these ends but more likely will lead to prototypes that do not scale but could provide valuable information and insight into the nature of the original concept. At the other extreme solutions that are applied to many different domains and embodied in an architecture that is successful at scaling and therefore that could be the basis of a business and be very profitable.

3.1. Developing Effective Prototypes

Research in Artificial Intelligence (AI) has reached a stage of maturity where many important products and services are being created. Such products often include advanced functionality that has been developed and demonstrated through a series of prototypes [26-28]. We refer to these advanced prototypes as "working prototypes" as distinct from simple mock-ups that only demonstrate a visual conception of the idea.

Working prototypes are highly useful in exploring the problem space for developing products eventually, because these prototypes are driven by the use of some special tailored technology that incorporates substantial AI techniques. Working prototypes that demonstrate special capabilities have even led to important breakthroughs in how users can achieve their tasks. Intelligent assistants were initially implemented as working prototypes and were rapidly improved based on user feedback before becoming mature products.

Although working prototypes serve an important purpose in helping transition AI research into product capabilities, developing such prototypes is not easy. AI prototypes often need to use a combination of novel AI tools to demonstrate their capabilities and therefore require careful synthesis of these tools. This synthesis process often requires experience with implementing related AI tools. However, such implementations are often not available and might differ significantly from the scale and efficiency in the final product [6,29-31]. For instance, state-of-the-art AI-based visual object recognition may report high accuracy but relies on a costly heavy-weight support of multiple image samples from the client for training an individual classifier before it can be deployed on the mobile camera. Such implementations may be impractical for real product use and considerable additional work may be necessary to scale the approach effectively enough for prototype or product use.

3.2. Scaling Strategies for AI Solutions

While it is often hoped that impactful AI solutions can be deployed with minimal resources, delivering scalable solutions requires careful consideration from the outset. First, scaling is both a technical and operational issue. The operational element centers around organizational capabilities, especially around data engineering [32,33]. The technical aspects include both the raw computational and distributed model training considerations, but also the software stack being used to deploy the solution, as well as the risk of on-time-in-full issues. The most common pitfall for organizations deploying AI is underestimating the extent to which the created models are imbedded into wider business systems and reporting mechanisms that depend on large volumes of data.

Second, finding a sweet spot in the context of the application is critical. The need for additional, targeted development may be minimal for some AI applications. Additionally, delivery partners will sometimes make this risk return consideration for you by offering a no-cure-no-pay model. For some kinds of established policy optimization problems, the bulk of investment will be in developing new heuristics or working solvers [34-36]. However, in most cases, going to deployment usually has a high resource and software development cost, and across plenty of applications it is

running experiments while kept on a short leash – just to implant self control – which will achieve institutional understanding of the return distribution.

Given the need for many AI solutions to be incorporated into wider systems for ever-changing decisions, the design of the solution often needs to aim for flexibility. This is only heightened in AI safety contexts, and is clearly already an aspect being considered.

3.3. Case Studies of Successful Implementations

The transformative potential of artificial intelligence (AI) technologies has been exemplified by solutions for a range of relevant global challenges. AI tools and systems have now amassed a record of successful implementation in diverse areas such as climate science, collective action, COVID-19 response, ecosystem management, economic development and disaster response, education, governance and credibility, health and medicine, law and public safety, information, language and security, natural resources and energy, poverty alleviation, social and environmental justice, sustainable agriculture, and wildlife conservation [16,37-40]. Moreover, advances in models and techniques such as natural language processing, visual understanding, speech recognition, or machine learning democratization and optimization, are rapidly extending the boundaries of existing AI solutions.

However, to take prototypes from small testbed contexts to widespread real-world impact is a daunting task. The barriers stem from the diverse high stakes problems we are considering, and the meager success of research on interest alignment in basic, applied, or any of the AI safety, transparency, fairness, responsibility, or accountability fields, applied to actual real-world problems [41-44]. In the service of sharing ideas, identifying pitfalls and clarifying pathways, in this section, we lay out a few of the case studies for publicly available AI solutions that have made the jump to impact, and the scaling pathways that worked for them. These case studies provide important lessons, not just in the idea-to-govern action pathways but also the impact and virtue incentives from both demand and supply sides without which demand-supply matching may not be as readily available. The learnings from these cases is not a canonization of the temporary, but rather an outline of pathways, actors, resources and key bridging points that allowed for a diverse set of AI prototypes to go from what might well be brief interludes of academic curiosity to chapters in the annals of engineering for the public good.

4. Policy Advocacy and Global Equity in AI Deployment

As the pace of technology disruption increases, it is imperative that more entities advocate for AI policy, development and deployment that protects and benefits the

most vulnerable in society both on a national and global level. There are policy decisions today related to the development and deployment of AI that have profound implications on creating disparity in who ultimately benefits from the billions in value that AI generates. Governments deploying AI on behalf of their constituents should be doing so ethically, but also making sure that its constituents, particularly those most disadvantaged, are both protected and educated so that it can hopefully elevate them to create equity. This requires both a reevaluation of education systems in light of the demands of the new economy and an increase in direct feedback mechanisms that let people communicate their needs to the corporations and governments creating the policy decisions guiding our everyday lives. The bottom-up feedback mechanisms are necessary because policy must evolve alongside testing and deploying AI; AI policy can no longer be a slow, iterative governmental process. Legislators must embrace a flexible, feedback-based approach to better ensure support of the system with the ultimate goal of protecting the rights of individuals.

Understanding AI Policy Frameworks

With everyone from governing bodies to elected officials creating their own AI principles guiding the deployment of AI, understanding these principles, frameworks and strategies is becoming imperative for the governing of AI policy [1,12]. The sheer number of different entities creating AI governance frameworks signifies a recognition that we as a group want to guide AI policy in ethical, responsible and equitable pathways. Look at the various entries in the table to see which principles resonate with you, your values, and the changes you want to see in the AI policy landscape. To ensure AI is developed for the benefit of all and taken all across globalization, it is critical that we as the creators of these new AI systems, policymakers and civil society members advocate for equity and access as additional AI principles. These could mean anything from increased access to free computing resources to decreasing the trade secret protections around algorithm sharing so that advantage does not simply tilt toward those with the ability to pay and gives back to the trend of personal and corporate accountability that suddenly seem to be losing prominence in society.

4.1. Understanding AI Policy Frameworks

AI systems are built and fine-tuned in a global landscape of cultural differences from established ethical perspectives. From the Westphalian concept of state sovereignty and global affairs as a zero-sum struggle for power, to Confucian principles of loyalty, propriety, and hierarchy that undergird most of the Asian political world; from a collective sense of news such as Ubuntu as reflected in the traditional African community, to the idea of mutualism that arises from Buddhist teachings; the differences and potentials converge on the real-world artificial intelligence solutions

adopted to address local problems, which should correspond to the cultural contexts in which these AI systems reside and operate. For the development and deployment of AI solutions, this calls for a feedback loop not only from the end-users to the politicians with respect to the particular nation-state that these economic and business interests represent, but also from AI experts to the AI policy institution makers and economists charged with producing those policies in order to ameliorate instances in which local approaches diverge.

The policy-making establishment is comprised of the political actors as well as the institutional environment of legislators, bureaucratic structures, and political culture, both formal and informal. Policy analysis uses scholarly tools and techniques of economics, sociology, psychology, systems analysis, operations research, and political science, to develop and evaluate arguments and proposals from which stakeholders will choose and that action will be based on. AI policy endeavors to integrate a patchwork of regulations from the few areas in which AI guidance has begun to take shape, from opt-in consent for experimental AI in healthcare, to civil service modernization and the government’s use of technology in support of democracy, to monitoring systems that prevent the impersonation of persons or departments in the political world.

4.2. Addressing Global Disparities in AI Access

Representatives of more than 60% of the world's population arrived at the conclusion that more AI solutions are needed for middle- and low-income countries. The implication is clear – A significant share of the world's population, with massive unmet needs, do not have thriving opportunities to access the benefits of modern AI. We call for collaborative and investment initiatives. In an open letter to the global AI community, a group of world-leading researchers wrote, "AI tales suggest realignments of existing scientific disciplines, often leading to closure of labs at big government-funded universities or disparate groups at big corporations" and argued for the creation of "self-sustainable self-owned labs which will take up the challenges of creating the large amounts of quality data needed for AI models which are capable of understanding languages other than English." Countries and regions with weak scientific and industrial capacity, responsible science and technology policies may require policies to ease disparities in AI lab creation, infrastructure and development funding.

Policy work on the global divide has, at least outside of economist-led initiatives, been missing a multilingual global and historical sense. In policy terms, two areas with stronger international action would be the implementation of pro-competition, consumer welfare-based policies and investments in R&D and digital infrastructure,

especially addressing local language deficiencies. Governments are very much needed in both areas, especially and more strongly supporting bottom-up initiatives breaking away from English dependency. In line with the focus on equitable AI access, the launch of a European Digital Commons and calls for a Global Code of Digital Cooperation call for fostering solidarity-building values in strategically important areas, like data and infrastructure, that are critical for the development of AI.

4.3. Ethical Considerations in AI Deployment

The Ethics Guidelines for Trustworthy AI emphasize that for AI to be trustworthy, it must be lawful, ethical, and robust. With respect to quality, trustworthy AI should be technically robust and reliable, and the deployment of AI solutions can be framed within the Compliance canonical that recognizes the core role of quality in the delivery of a trustworthy AI solution. Additionally, ethical assessments guarantee that data is used in a moral manner. Thus, ethics deals with the "data question," addressing what ought to be done, i.e., what could possibly justify the AI solution if something goes wrong. Importantly, ethical considerations do not only correspond to the solution itself, but also to how it is developed, deployed, and put to use. AI policy regulation can ensure accountability, compliance, auditability, and on-going sanctioning, which are paramount within ethics. This includes establishing ethical, moral, and social standards that provide the perimeters within which AI solutions ought to be developed.

In this respect, ethical reflection addresses the human consequences of AI development throughout its lifecycle, investing the deployed AI solution with a component empowered to determine whether a responsible impact is being delivered. In fact, there is broad consensus regarding the paramount importance of human agency and oversight through the entire life cycle of AI technology. Yet, responsibility for irresponsible impact does not only correspond to developers of AI solutions; users who deploy AI solutions ought to guarantee that these are designed, developed, and used in an ethically correct way. Ethics, therefore, addresses the questions of who decides whether a solution is responsible or not and how such decisions are made.

5. Challenges in Translating AI Research

Translating AI research into impactful solutions that affect real people and communities throughout the world is not an easy task, and many research projects stay in the prototype stage without any further development. There are a variety of factors that can hamper research-to-solution translation in AI. While the barriers can be quite different from project to project, it is helpful to identify some of the more common challenges that occur often and look at how various translational AI research centers around the world have addressed—or are addressing—these key challenges. For example, as is common in many interdisciplinary academic fields, AI research can sit

at the intersection of diverse academic domains, and effectively engaging people with different expertise to foster inclusivity and build upon each other's knowledge and skillsets is not always an easy task. A common challenge in translating theoretical AI research into practice is navigating the landscape of rules, regulations, equipment requirements, and grant funding opportunities. Not only can this landscape be complex, there are often unintended obstacles that impact research participants, often in localized, specific ways.

Solving these issues requires communicative collaboration between the lab and the field—all team members need to express their needs and concerns, and want to learn about the needs of others in order to create AI solutions and deploy them in a way that is beneficial for the people these solutions are made for. However, this is often not easy to do in practice. It may seem simple to form an interdisciplinary project team, but building trust and a single vision of success between deeply knowledgeable people from different backgrounds, with each background lending important skills to the project, can take a significant amount of time. Robust AI solutions often need to be able to robustly work for a wide variety of inputs, typically have low operational costs, and work with little maintenance. Depending on the project, they may need to have some capacity for complying with user feedback as well.

5.1. Barriers to Cross-disciplinary Collaboration

A critical challenge in the scaling of AI-based solutions is bridging the gap between the translation and adoption of novel methods of AI and the advances and availability of well-validated AI processes and tools that are transferrable. Indeed, most of the major bottlenecks in translating AI models into solutions are interdisciplinary ones, for example, inter-ministerial coordination when focusing on public service and policy applications, practical sector-specific knowledge when aiming at business and ethical implications in the private sector, and a myriad of nuances sensitive to cultural diversity, historical contexts, governance, social dynamics, and enforcement regulations when applying to LDCs and emerging economies.

Establishing sustained partnerships bridging different knowledge domains and translated expertise demands identifying mutually beneficial structures and spaces, e.g., within the public sector and industry, shaping embedded incentives and resources that support collaborative projects bringing measurable long-term action impacts. In fact, cross-disciplinary collaborations that unify AI researchers with social and behavioral scientists are essential to supporting the actual uptake and use of AI methods into policy and sociotechnical applications. Indeed, whilst AI research has seen booming growth in the past decade due to the extreme proliferation of publicly available data and trained models in natural language processing, computer vision, etc.,

the value of translating and adapting these advances to locally attuned solutions that accelerate public policy transitions bringing social impact will continue to overshadow more scientific explorations focused on live demonstrations and hypothetical applications.

5.2. Technical Challenges in Scaling AI Solutions

The vast potential of applied AI research can only be realized when the models move from prototype level to system level that is capable of handling millions of queries, generating machine-supplied data in large quantities and with sufficient accuracy for downstream tasks that are critical for business, making company-wide decisions based on AI models etc. But getting it to production is a hard problem. One of the problems is there is often limited resources available for massive scaling. Often, the models trained are resources hungry and need fine-tuning to run on edge with punchy user experiences. Another problem is once such resources are available, the goal posts for what it means to scale keeps changing. Sometimes the solution needs to be production-ready for low-cost, real-time inference. For instance, NLP tasks like paraphrasing or grammar checking need to be done fast and cheap for large-scale applications. It is easy to reach a pipeline-level solution where models can be parleyed.

But deploying it into production is less trivial. Such systems need to be automated to serve millions of users. They need to support topics like creating models for user tasks that need no or few examples, deciding user intent for large number of sentences in real-time, invoking and serving machine-supplied paraphrases and keeping quality checks on various involved models including quality of paraphrases through several metrics such as semantic similarity and lexical similarity. Similarly, sentiment analysis tasks need to be done real-time and cheaply especially in marketing and brand monitoring scenarios. But when cyber scientists for a large firm at one time took to these tasks, the accuracy of models was far from production-ready.

5.3. Navigating Policy Landscapes

Policies addressing the challenges AI is being developed to mitigate — such as unemployment, inequality, and climate change — may affect the components of the systems utilizing AI. In many of these areas data is deeply sensitive, not least because these problems are often aggravated by structural inequalities and histories of discrimination. Anomalies and outliers are a higher ratio of these populations, and therefore can skew results, and reinforce existing inequalities if care is not taken. Using AI in these sectors is further complicated by policy associated with the sectors in question. The question-cascading data and bias concerns associated with public deployment of private technology in health care are extremely complicated, and have

explicit precedent in how specific tools are used and what ethical framework underlies these areas of deployment. For example, AI deployed for triaging in radiology would be subject to clinical verification, validation, and monitoring, similar to clinical decision support tools, an area where many companies writing these systems have run aground with various forms of enforcement from various policy and regulatory agencies. Other algorithms in radiology are viewed more circumspectly, as data differs in how widely it is publicly accessible or validated. This patchwork of regulatory environments affects both private deployment of private technology in these areas as well as widely used tools.

Transparency has become a touchstone of responsible innovation, and one mode of securing greater transparency is what is being termed third-party advocacy. Advocacy by civil society organizations such as scientists and auditors accomplishes widely acknowledged goals such as leading AI researchers and companies to work on many problems that matter to them, incentivizing submission of these tools to regulatory processes, and ensuring assignment of liability in the case of failure. Co-regulatory models have often been discussed in detail.

6. Future Directions

AI Research has made remarkable strides that have the potential to improve many facets of our daily life. However, as AI models grow larger and require vast technical expertise, scaling them to solve high-impact problems is challenging. As more organizations develop expertise in developing language models and more public discourse centers on the societal implications of AI technology, we discuss some emerging trends in AI research that can help us identify the most impactful paths forward. These emerging trends can enhance our ability to generate workplace solutions by forecasting directions for AI Research that can facilitate more collaborations between academic research and industry, as well as suggesting innovative policy support that can help to align AI Research towards impactful areas of product development. The rapid emergence of large generative models have surprised and inspired researchers and practitioners in many disciplines. These efforts have also sung the praises of accessibility - making models open source, inexpensive to use, and readily available to developers. As large language models reach incredible benchmarks just days after their release, other deep learning areas are starting to see immense models that are addressing multiple downstream tasks, such as vision-language representation learning and multimodal generative models and the development of open training datasets as a parallel effort to boost academic research. At the same time, new techniques are advancing the art of open-sourcing these models, such as hyperparameter pruning, distillation, quantization, and other techniques for efficient inference. As we continue to experience the potential of AI and its future uses in

multiple areas, the call for federal action in AI oversight is also at the forefront in other areas, such as advancing equity and fairness in policy priorities and shaping the role of working people in an economy increasingly transformed by automation.

6.1. Emerging Trends in AI Research

What is going to happen in AI research? The aim of this chapter is to outline some answers based on recent trends as well as the expert opinions we discussed in the previous chapters. While not the ultimate answer, we can at least respond to the question regarding the future direction of AI research that is derived from 2 small and more specific sub-questions such as: what are the current research directions in AI that are underexploited? What are novel technology verticals within which to innovate AI research for the greatest potential positive impact in the long run, even possibly going beyond the functionality itself? In this chapter, we explore the future direction of AI research that answers to these 2 questions.

Modern work in AI has drawn from and contributed to a variety of fields like information theory, cognitive science, neuroscience, computational linguistics, control theory, embedded systems, evolutionary biology, philosophy, social resonance to mention a few. That said, deep learning has been a phenomenal success with new developments showing promising results almost daily. This has raised new questions about the fundamental limitations of the approach and has forced researchers to more directly address broader AI goals, like robustness, sample efficiency, transparency, controllability, and learning in more naturalistic domains. Reinforcement learning, generative models, temporal summarization, and autonomous discovery of discovery procedures are examples of areas where emerging research trends appear to contribute more directly to addressing these challenges. These lines of inquiry are exemplary de facto research programs because they not only address instrumental AI goals of importance for powerful AI but also are associated with the prospect of substantive near-term applications. For example, RL and generative models have been applied with some success.

6.2. Potential for New Collaborations

The key for creating new theory-to-practice bridges is to develop ideas, methods and systems which are both immensely powerful, and flexible to fit into a diversity of practical applications in cooperation with domain experts. Collaborations between social scientists with domain knowledge, and technical experts from the technosphere are rare, and often only take place after the technical researchers have already “discovered” the social domain without much expertise or grounding. However, while many methods have been adapted to the social science domain at a superficial level,

there have been few attempts to change the actual state of a social science field, or contribute new insights. This is largely due to the many gaps or unresolved issues in domain expertise or collaboration infrastructure between domains.

We believe the time is now for truly earnest starting points for interdisciplinary collaboration. The timescale for transitioning from a successful theory to foundational solution is similarly becoming shorter and shorter. While a decade ago deep learning experts often posed new systems which used the last stage or a part of an older system, now training multi-stage solution architectures from proposals in a foundational area such as transformers, and specializing on application areas is en vogue within the AI community. As domain adaption becomes not just possible, but also easy, it will be the right time for social science domain experts to collaborate with the technosphere in earnest.

6.3. Innovative Policy Approaches

Creative AI will have profound implications for the roles that technology, policy, and business play in determining the trajectory of a diverse and vibrant creative sector. As AI systems become increasingly integrated and capable — with the ability to generate new content, tools, and even business ideas — what role should public policy play? Governments have long played an essential role in ensuring vibrant and diverse creative industries, by supporting artistic expression and the tools that creators use. But with advances in generative machine learning, the technology world increasingly offers creators powerful new tools that might upend some traditional government functions, while in other areas it seems unlikely to deliver for artists, or society more broadly, without more public support. The need for creative human actors may remain or even grow, but over time, the roles played by creators, technology, and government, and the relationship between them, will shift. As these changes happen, and the technology challenges traditions and assumptions held by many in the creative sector, innovation policy will need to maintain robust support for artistic expression, while embracing aspects of Creative AI. Creative AI and the machine learning research behind it have implications for technology and media policy that go broad and deep.

Key questions about future work in this area include how best to help new people enter a variety of creative fields; how can AI be harnessed to enable and leverage those people; and how can AI assist established risk-takers to produce and distribute their work? With generative machine learning capable of creating raw materials for the development of culture, we must also ensure the inputs to this AI are sourced ethically. AI raises issues of authorship and ownership, both for individuals developing personal style and technical chops, and larger exceptional artistic bodies, but also for large

companies generating their own data proxies, and then contracting talent to create with minimal input.

7. Conclusion

In the past, researchers have focused on demonstrating new ideas and showing their initial superiority for relatively small benchmarks. Now, the AI community has reached a point where ideas are so plentiful that attention has shifted to specific and well-defined tasks, for which companies and the community are working on better solutions. Publicly available data for benchmarking is immensely useful and companies have spent substantial resources collecting it. Similarly, businesses have started to release their solutions to texts, APIs, downloadables, and platforms for the community to use. In this effort, some have prioritized knowledge, performance, stability, and safety, while others have favored genericity and ease of use, or rapid iteration over all the above.

AI groups in academia need to transition from building better proofs-of-concept of ideas to developing all-things-considered solutions. Research will not only achieve greater real-world impact with this approach, but it will also become more helpful to the community. Companies are willing to favor decisions that get more attention from the community, usually the decisions whose consequences they can assess better, either because they are further detached from the core problem, which becomes more expensive to measure, or because the corporate resources diverging from the core business are significantly smaller. Fitting into a community's roadmap is a necessity. Unfortunately, solving a problem better does not guarantee it will capture attention and immediate rewards.

References

[1] Durai S, Manoharan G, Ashtikar SP. Harnessing artificial intelligence: Pioneering sustainable solutions for a greener future. InSocial and ethical implications of AI in finance for sustainability 2024 (pp. 89-117). IGI Global Scientific Publishing.

[2] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.

[3] Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, Abdullah S. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomedical Materials & Devices. 2023 Sep;1(2):731-8.

[4] Ozkan-Okay M, Akin E, Aslan Ö, Kosunalp S, Iliev T, Stoyanov I, Beloev I. A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. IEEe Access. 2024 Jan 18;12:12229-56.

[5] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.

[6] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.

[7] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.

[8] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.

[9] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.

[10] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.

[11] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.

[12] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.

[13] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.

[14] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.

[15] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.

[16] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.

[17] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.

[18] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.
[19] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.
[20] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.
[21] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.
[22] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.
[23] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.
[24] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[25] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[26] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[27] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[28] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[29] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[30] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[31] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.
[32] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.
[33] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.

[34] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.
[35] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.
[36] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.
[37] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.
[38] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.
[39] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.
[40] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.
[41] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.
[42] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.
[43] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.
[44] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.

Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-745-1

DeepScience
Open Access Books

Chapter 13: Toward Generalizable, Responsible, and Adaptive Artificial Intelligence

Samit Shivadekar
University of Maryland Baltimore County and Research Associate at Center for Accelerated Real Time Analytics (CARTA) UMBC, United States

1. Introduction

Research in Artificial Intelligence and Machine Learning has made remarkable advances, achieving super-human results in some specialized tasks. Multimodal and general-purpose models have attracted particular attention, creating popular excitement and intense anxiety around the prospect of Artificial General Intelligence. However, most of these systems remain brittle, fragile, and opaque — and worse, highly susceptible to malfunctions that can lead to irresponsible and harmful consequences across a spectrum of domains [1-3]. Issues stemming from cybersecurity vulnerabilities, societal biases, and environmental footprints, amongst others, must be addressed decisively before we achieve the goal of truly Artificially General and Responsible Intelligence.

There is increasing agreement across the research community, industry partners, and policymakers, regarding the need to make general-purpose AI systems more trustworthy, robust, interpretable, and responsive to guidance and feedback [2,4,5]. These properties may not only help mitigate, but also eliminate undesirable and dangerous behaviors exhibited by these systems. However, while there may be some consensus on the type of normative safeguards that must be put in place, the underlying technological challenges are less clear. More particularly, because of the enormous capabilities of these systems, addressing each of these challenges is non-trivial, and achieving the desired improvements may not be straightforward. In this paper, we identify important technological challenges related to general-purpose AI, and offer inspiring thoughts that are motivated by our experiences working to create responsible and adaptive AI. We hope these thoughts will serve to initiate a dialog

towards framing a research agenda that will enable our AI systems to be more trustworthy, user-interactive, and aligned with human preferences.

2. Lifelong Learning, Domain Adaptation, and Resilience

It is impossible to know a priori all the things AI should know, what the environment looks like, what risks the agent is going to face, what other agents are interacting with it, and what particular species of agents that AI is supposed to observe, learn, and reason about. Thus, it is the version of AI we are going to build today that needs to start learning new skills, new knowledge, adding and adapting capabilities incrementally, without excessive resources, given the challenges of simulated and real-world deployment [6-8]. Therefore, making these AI systems generalizable, adaptive, resilient, and responsible is the most tantalizing, exciting, and significant challenge we should aim for. Adaptive learning is the AI agent's capacity to learn from the environment, or other agents when receiving feedback or reward functions, and change its behavior accordingly. More generally, it reflects not only the ability to improve but to acquire new skills and behaviors. Infants and children in particular have generalized learning methods for recognizing increasingly complex structures in the world, alone or from parental support [9,10]. They build increasingly sophisticated and knowledgeable domain theories and models that help them process environmental information. They have cognitive development trajectories that show how junior learners can perform at a high level in many domains in which adults are novices, while adults are experts for the domains where children have the least domain knowledge. These trajectories suggest how to design developmentally friendly learning interfaces, scaffolding methods to reduce the likelihood of catastrophic failures, and also learning difficulties designed to challenge domain generalization powers and domain-specific knowledge. The combination of general knowledge about the world, prior knowledge about specific domains, and learning scaffolding tools provided by parents and peers leads children to construct sophisticated hierarchical, compositional, and domain-specific knowledge structures in a way that avoids the pitfalls of overfitting in off-policy learning.

2.1. Conceptual Framework for Lifelong Learning

Artificial Intelligence (AI) has rapidly influenced both shaping and solving some of the challenging problems faced by humanity. While algorithms can now perform specific tasks at or above human capabilities, they are trained artificially on labeled data in a disjoint fashion, primarily using supervised learning. These algorithms are able to generalize well on them according to some performance metrics. However, in most real-world applications, these systems can be found stuck in a distributional regime, as either the distribution of data or performance metric for a given task can shift, posing

novel challenges to AI systems. Are such algorithms capable of generalizing or adapting to these shifts? The problem of re-training the underlying models using newly collected samples from the target distribution is expensive, cumbersome, and not always possible, e.g. the worst-case scenarios where there are no labeled target samples available. Thus, the central ideas behind the concepts of domain adaptation or domain generalization techniques have been to devise AI algorithms that can adapt to or generalize on tasks with unseen distributions.

The general paradigm of lifelong learning, also known as continual learning or incremental learning, seeks to study how an AI system can continually learn new tasks while already being provided different shift experiences and also minimize the negative effects of catastrophic forgetting [11-13]. Lifelong learning hence aims to develop competence while adapting in a flexible, efficient, and data-driven manner to new or modified distributions, directly affecting the resilience of the system. While the concept of domain adaptation looks at the transfer of knowledge from a set of labeled source distributions to an unseen target distribution, the underlying problem of lifelong learning is a temporal one, looking at the effects of being exposed to the unlabeled trajectory of such shifting distributions over incremental periods of time on the learning process and updated model. Lifelong learning is apparent throughout human evolution and learning but is still far from being realized in machines.

2.2. Techniques for Domain Adaptation

Domain adaptation is an effective technique in lifelong learning to address distribution shifts from one task to a related task by bridging the learning from an earlier task to the later one. In domain adaptation, shared knowledge between domains in source and target tasks, as well as domain-specific information, are exploited to transfer what has been learned in a specific domain to an unseen but related domain. Knowledge transfer between related domains can substantially improve performance in the target domain, especially when labeled target domain data is scarce [2,14-17]. In general, domain adaptation methods specifically model and reason about domain relationships and design learning objective functions to capture the differences in knowledge properties of shared and private components among different domain models. Broadly speaking, most domain adaptation methods exploit sources of information in task relationships via transfer learning, probabilistic models, metric-based learning, or generative models, among others. Transfer learning, especially its early formulation in embedding representation learning, is perhaps one of the earliest approaches to domain adaptation, where input data is transformed into another space and the domain relationship is modeled in this space. Transfer learning assumes the domain mapping can be learned through supervised data labeled with their respective domains, which is often impractical [9,18-21]. To overcome this limitation, domain adaptation aims to learn

domain-invariant representations through the use of information from both domains. Domain adaptation incorporates shared knowledge between the source and the target domains and augments the transfer learning methodology under conditions when the two domains are related.

2.3. Building Resilience in AI Systems

Towards Adaptive and General Purpose AI beyond Labeled Closed World Tasks Lifelong Learning, Domain Adaptation, and Resilience For any AI system to be applied in a wide array of tasks in the physical world, it needs to be capable of rapidly adapting to knowledge and purpose in the presence of scarce or no external guidance or supervision. Our ultimate vision for AI systems is that they will be computational assistants for our work and everyday living, just as our use of computation to augment reasoning, recall, skill acquisition, management of everyday living are aids for our lives. In a very loose analogy land and ocean species react and adapt to sudden shifts in their biome, such as earthquakes or hurricanes [22,23]. It is adaptation to these unexpected changes which is crucial for the long term continuous existence of all species. A parallel in the AI domain of supervised learning is a major shift in sample distribution, resulting in major degradation in classification accuracy. Resilience, in the context of Lifelong Learning and Domain Adaptation, is the ability of an AI system to incrementally self improve as more and more applications are processed by the AI module. To understand resilient learning, it is important to understand curriculum learning. In Curriculum Learning, the learner is trained on a sequence of labeled tasks of increasing difficulty – the easy first principle – suggested from the way humans acquire skills. The starting and intermediate exercises in Curriculum Learning are carefully chosen so that gradients are informative. As long as the learner does not forget previously learned tasks, it will eventually have perfect performance on the entire sequence. In the concept of resilience for AI systems, the entire process is self supervised, may proceed totally without labels, or be done in coordination with potential usage experts.

2.4. Case Studies in Lifelong Learning

The examples in this section demonstrate the plethora of applications and capabilities where lifelong learning can greatly enhance the potential of existing systems, while also serving as a bridge during the evolution from narrow- to general- AI. In translation, domain adaptation facilitates the generalization of models outside the training domain, but achieving natural language processing in many languages simultaneously is an enormous task; hence a single language through domain adaptation makes utmost sense [24-26]. The same is true for many modalities—learning from the inconveniently sized constrained data available reveals whether and

how much it helps if leaned and grown in lesser languages—what's the silver lining. It hardens the model through few-shot active learning; not managing to wait just a few more seconds until the completion of the graph traversal gets the screenshot that is eventually used during the work stream that shows how they can help enhance the process.

An open challenge is the unavailability of sufficient and diverse training data. Frequently, only a few examples are available, even in the target domain. For many tasks and domains, it is impractical to annotate and use tremendous quantities of training data due to cost and time constraints that are unacceptable. Some tasks that machine learning can perform simply cannot be modeled with that compromise. Equipment hosting and processing the algorithm must therefore learn to adapt without a performance drop during deployment. Here we show some case studies illustrating the use of lifelong learning strategies as well as techniques for domain adaptation to help enable machines operate on edge-type tasks on emerging devices. Examples are considered from cross-lingual and multimodal applications, robotics autonomy, risk-life operation for disaster-relief operations, and other relatable ones.

3. Human-AI Co-Learning and Feedback Loops

Indeed, learning is a two-way street where knowledge, skills and expertise flow in both directions; without a doubt, this is even more true when considering the educational collaboration between an intelligent agent and a human user. An AI system learns from the human agent providing it feedback in different forms, and is therefore co-creating the information it possesses; but at the same time, the user as a biological intelligence is learning from the intelligent agent, adapting its behavior to the feedback it receives from it, improving the effectiveness of the collaboration [27,28]. Note that humans collaborate and solve problems together; AI systems collaborate and augment human intelligence to resolve joint goals. Collaborations and joint tasks between humans and AI systems differ in natures, and therefore they necessitate different designs and implementations.

Analyzing the feedback stream is fundamental: in Human-AI collaboration, the main task is to generate feedback-related representations. The more "introspective" the AI system is, the more it is able to provide specific information about its inner states, the more useful, personalized, and effective the feedback will be for the user. We therefore can put forward the Basic Co-Learning Proposition: empathizing the inner workings of the AI agent is a key feature of the user-AI interaction; producing information and empathy is essential for user learning and skill improvement. Agent internal states therefore need to be shared during the collaboration; psychological investment for AI design would pay back for the entire community. These are qualitative arguments that

show that co-learning relationships are intricate and sensitive: while pushing forward the AI and the human in a complementary way toward the solution of the problem is necessary, the balance between AI and human disparateness should be indeed finely tuned [19,29-31].

There are qualitative psychological and cognitive reasons that support the necessity of a fine-grained balance between human and AI learning mechanisms, and of related adaptive scenarios based on different collaboration stages. By "adaptiveness" we mean here a switch of roles and related styles of collaboration, according to the phases of the project: collaborating with compassion, flexibility and differentiation with each other, and being aware of both AI and human profiles is essential to raise the quality of results.

3.1. Understanding Co-Learning Dynamics

Co-learning and co-evolving effectively implies that humans and AI learn from each other, enhancing their respective learning and learning capacity, in a positive feedback loop. For the AI component, this enables improved learning capability with less data and improved generalization. While humans are already adept at providing feedback for AI in the task I can do, and the AI technology is increasingly reliable in context if not the content of the output, currently we cannot easily teach the AI to perform a task outside of the mechanisms already provided, for example using natural language, even for tasks that are easily learned by a child. We envision future AI that can accept extended feedback beyond current simple feedback or correction, and that learns efficiently from the feedback, taking it into account.

For humans, effective co-learning with AI should be an engaging experience that is relatively easy and does not require too much invested effort. Part of this co-design is that humans do not need to understand machine learning in detail, but we do believe that a basic understanding of how AI learns might make it easier for the human to design appropriate feedback mechanisms, especially at the earlier stages of learning when the AI is still adapting to the human's requirements. Good models of human cognitive biases could help ameliorate the effects of poor design of AI tools that pre-exist, or new tools that have been poorly adapted to the user's actual learning pattern. We see co-evolution as a partnership over time which is currently hampered by the gap in understanding of how AI learns given little feedback in teaching from humans.

3.2. Designing Effective Feedback Mechanisms

Effective human-AI collaboration relies on an intricate co-learning relationship, and therefore we need to design effective AI-enabled feedback mechanisms to facilitate this process. One common method is through reward signals - in interactive

reinforcement learning setups, the human provides feedback by delivering reward signals to the AI and the agent uses these signals to learn a reward model and use it to further optimize its policy performance [32,33]. The overall aim is to enable humans to communicate their desires, decisions, and intent to the AI efficiently and improve AI learning.

A key aspect of these feedback mechanisms is their fidelity, or, how accurately these signals convey all of the underlying human intent. Humans, while providing feedback, might not be experts, and engaging all of their efforts might not be viable for all settings, causing the feedback to be noisy or inaccurate at times. This typically leads the AI, while optimizing its policy based on the imprecise or incorrect signals, to not learn optimal behaviors. One approach to alleviate this problem is to allow the AI to ask clarifying questions, and choices, while trying to reduce the need for humans to give feedback [34-36]. When expressed as a generative feedback model, researchers have proposed the approach of Conversational IRL, where the agent can engage humans in a dialogue to confirm or clarify its predictions. The dialogue can confirm which parts of the prediction are likely to be wrong, and can also clarify ambiguous predictions with two or more likely explanations.

3.3. Evaluating Human-AI Collaboration

Before studying how to design human-AI interaction to achieve effective human-AI collaboration, we must first answer the question of how to evaluate whether such collaboration is effective. This is complex because the success of the collaboration is inherently subjective. In purely artificial agent AI research, task performance is usually treated as the sole evaluation criterion. Such task performance may be measured using objective means, such as the time taken to complete a task or the accuracy of the produced solution. However, much of pure-AI agent research focuses on the stated goal of the AI agent itself, either deterministic or stochastic. This necessitates some physical world metric to capture the essence of the task as viewed by the human designer of the agent. In the case of designing systems to directly interact with humans, success is considered as instances of task performance, combined with user satisfaction and trust in the system. In general terms, first-hand user experience during and after the interaction becomes a relevant criterion of success.

In the Human-AI co-learning framework, there is a different objective, the dual task of designing effective interaction mechanisms and communication displays that allow user understanding of AI uncertainty while at the same time improving the recognition and prediction capabilities of the AI [37-40]. This objective adds new qualitative criteria to measure the effectiveness of collaboration; for example, the quality of explanations provided by the AI, the transparency of communication, and how the

interaction, with or without the use of communication aids, affects the burden of communication on interactions. We assert that both pure task performance measures and criteria based on understanding and user experience are necessary to assess the success of a collaborative research approach. After all, a human-AI combination can only succeed if both agents perform better together than each one would have done alone.

3.4. Challenges in Co-Learning Environments

The dynamics associated with the feedback loops and co-learning will depend on implementation details and how the task is structured. Sometimes the human provides explicit feedback, and sometimes the AI presents intermediate work for review. The asynchronous nature of tasks can also affect system dynamics. In some use cases, the human will furnish examples over an extended period, and in other scenarios, the person may work concurrently with the AI, leading to a more team-like relationship. Integrating human feedback needs to empower and inspire humans to direct the learning process [41-42]. With weak feedback, the AI may overfit the response to the idiosyncratic human preference. Combining the weak critiquing feedback of the human rater with a larger dataset can help generalization or transfer the learned reward function more effectively. Different types of co-learning systems would face this problem in different ways, and some co-learning systems may benefit from this risk to express more focused understandings of persons. However, the design of co-learning systems is complex.

Co-learning environments present many challenges in addition to those faced by standalone AI systems. A systems perspective is necessary to examine such co-learning architectures. The AI benefits from the human input, but what is the payoff to the human? Humans should not be misled by the AI. Humans should be able to affect the process of learning or reward function design at one or more levels of granularity and control. Humans do not stay stationary over the long time scale of AI learning, particularly with interactive systems that engage in action-selection tasks. Humans may come and go or change in focus or even policy. Designing AI that learns from humans in the long term is related to modeling human behavior, although the complexity of this problem is also shared with human-centric interactive AI.

4. The Role of AI in Achieving SDGs and Climate-Health Equity

Artificial intelligence (AI) has transformed how we access technology and information and how we connect with one another and the planet. The last decade has seen unprecedented growth in AI investments, products, and applications, as well as an increasing commitment to diversity, equity, and justice. AI has delivered a myriad of applications in a diverse number of fields including education, employment, health,

energy, ethics, transportation, and climate change. Notably, addressing the challenge of climate change and climate-induced inequities in health, well-being and access to resources will require a core understanding of how technology coalesces with policy and practice, bridging the social, behavioral, technical, and environmental sciences to design and deploy AI systems that are accountable, responsible, equitable, and link agency to impact.

The challenges related to the COVID-19 pandemic exposed long-standing inequities that can only be resolved through an integrated approach that lifts people out of structural inequities in access to health, housing, education, and nutrition that ultimately create healthy societies built on partnerships. Further work is needed to reduce the burden on professionals, engage with domain experts, and seek specific applications that can contribute tangibly toward people's lives, the environment, and how we deepen our understanding of each other, and social constructs around trust and responsibility. In doing so, we believe that researchers, policymakers, and practitioners can drive this collective vision while better preparing current and future innovators to grapple with the questions of ethics, responsibility, and diversity in AI to ensure these challenges are met head-on to build a brighter tomorrow.

4.1. AI Applications in Sustainable Development Goals

With multiple intergovernmental institutions promoting global health, climate, and development initiatives and universal norms, we must elicit the values and capabilities of AI so that all have just and equitable access to AI development processes while benefiting from the use of AI. The world is at a leadership moment as countries and diverse entities across geographies dialogue to define the values, equity, and interest-driven joint global investment and collaboration mechanisms needed to advance an equitable and universally beneficial AI.

AI has become part of our everyday conversations and tools. More than ever before, we can imagine and benefit from the appropriate use of AI to extract signals from big data, optimize and coordinate resources and services at scale, and use AI for AI to rapidly innovate and reduce costs in all industries. We can apply AI to accelerate progress towards the SDGs in novel ways by leveraging deep learning and reinforcement learning, as examples. AI can be integrated with sensor networks, simulation models, and conversational agents to spur progress in sectors tied to a number of SDGs. AI is already used in myriad applications of targeting, prediction, monitoring, deployment, and measurement and evaluation in development and health, climate, and humanitarian assistance domains. In low-resource settings, small, low-cost, computationally efficient AI chips are being deployed on mobile or handheld devices for predictive analytics, task coordination, and real-time decision support. The

increasing scale and richness of data drive momentum in AI capabilities and applications.

Climate change is inextricably linked with health and development across these interconnected domains. Creating more value for people requires sustained innovation that brings together synthetic and human intelligence, allowing for algorithms and logic to augment reasoning, creativity, and relationship-building in small and big decisions and operations. AI can be applied to spur innovation in decision-making systems. AI is being combined with models for decentralized intervention strategies to support mobile agents, offer unique learning opportunities of data-driven behavior modeling, and incentivize simulated agents under adaptive learning.

4.2. Climate Change Mitigation Strategies

Implementing renewable power systems and establishing zero-carbon buildings are two major strides toward realizing a zero-carbon tomorrow. Available technologies can get us fairly close to a zero-carbon tomorrow. Carbon emissions as a byproduct of fossil fuel consumption in energy generation can be decarbonized by replacing fossil fuel-based electric generation technologies with geothermal, solar photovoltaic, concentrating solar power, nuclear, wind, hydroelectric, and sustainable biomass-based generation technologies. These sustainable technologies combined account for about 50% of global electric generation. Additionally, with burgeoning improvements in energy transmission, storage, and transfer technologies and algorithms, the wide adoption of electric vehicles, which will replace gasoline and fossil fuel-based coal burning in transportation, can further reduce the electric generation-related carbon emission problem. Another type of decarbonization employs the use of carbon capture, utilization, and storage technologies. It is expected that by 2030, available electric generation and carbon capture technologies can reduce global carbon emissions by 15%.

Making buildings zero-carbon require decarbonizing the electric power demanded for heat exchange by utilizing energy-efficient technologies to increase the current minimum energy-saving standards established for buildings. These energy-efficient technologies include thermal insulation of external walls, ceiling, and roofing; double/four-pane window glasses; thermal-heat exchanging vents; and state-of-the-art HVAC technologies that consume renewable electric power. Powering all renewable-enabled zero-carbon buildings will mitigate about 20% of global carbon emissions. By 2050, about 80% of buildings can become zero-carbon. The remaining dirty, non-renewable building sector of the economy can compensate by employing carbon capture technologies.

4.3. Health Equity and AI Interventions

Technology can help shape healthcare systems toward a more equitable future, and AI is increasingly been hailed as a critical enabler. Despite this promise, there is limited evidence that AI makes healthcare systems and processes more equitable. In this domain, health equity refers to the fair distribution of medical services and interventions, as well as their outcomes, with the goal of eliminating ethnic, geographical, gender, economic, and other disparities. For over 40 years, the fundamental principles of AI—efficiency, accuracy, generalization, and objectivity—have been challenged by sociolinguistic and anthropological critiques focusing on the socio-contextual and historical aspects of AI and on the risks of discrimination in algorithm design, validation, and application. In healthcare settings, the Joint Commission calls for the provision of effective, equitable, sequenced, and timely care for all surgical patients, with work directed at the mitigation of existing health disparities and inequities based on ethnicity, culture, gender and sexual orientation identity, and economic disadvantage.

Most of the earliest discussions of bias and discrimination in AI have drawn on the moral and ethical philosophy of epistemic injustice. More recently, and given the scale of AI interventions in medicine, calls for accountability have become increasingly insistent. AI researchers and developers, as well as those applying AI methods to the health domain, have been encouraged to move beyond simple claims of fairness linked to algorithmic performance, and begin to consider the impact and implications of deploying such models in clinical contexts. This focus on data, training, model evaluation, and real-world deployment of predictive AI returns us to the issue of health equity, and indeed extends and enriches these earlier conceptual arguments into a more pragmatic framework for AI interventions in health and medicine. Call for and prioritize action are welcome additions to the normative and ethical algorithmic justice conversations in machine learning.

4.4. Ethical Considerations in AI for SDGs

The race to develop innovative artificial intelligence (AI) solutions for solving the world's most pressing challenges has begun to catalyze a new layer in the global tech-led public sector engagement model where initiatives are shared and scaled in part to display the national prowess of soft power. National states increasingly direct their scientific and technological resources toward meeting humanitarian needs in low and middle income countries, as a means to fortify bilateral relations, define trade partnerships and mutually envisioned futures. The digital development approach, focusing on locally appropriate technology anchored to SUFs, as the basis for exploring a country's potential and preparing for new realities is critical to ensuring

that AI solutions are adaptive, sustainable and just. Globally, AI policy discussions have driven the rapid development of ethical guidelines to help inform and shape AI applications and tech-led funding decisions, which both avoid or mitigate any potential harms. Applied AI decision-making involves managing risk and prioritizing guidance in four key impacted areas, starting with the protection of human rights and civil liberties aligned to fundamental ethical and moral principles, ensuring diversity, equity and inclusion in the data, algorithms and the intended beneficiaries reflecting diverse stakeholder perspectives, as well as their continuous engagement throughout the life cycle. This is particularly important in the design recommendations and evaluations, so that these reduce or avoid any existing, implicit or abstract bias or discrimination. Second, public trust and government accountability in democratic systems should also be established through risk identification and mitigation in integrated decision making, guidance provision, outcome transparency, model interpretability, evidence-based research, and due process when AI decision-making occurs. Third, industry is called to address and avoid operational opacity along the digital supply chain involving the sharing or externalization of services and decision-making. Fourth, in response to climate impact avoidance, environmentally sustainable principles should be followed that also minimize the carbon footprint of the infrastructure and tools that support AI visualization, analysis, modeling and conclusion.

5. Future Directions in AI Research

This essay, broadly, argues for grounding and shifting the perspective in AI Research towards a focus on social and ecological adaptation. We see critique and pushback from a broad swath of academics, practitioners, educators and students of the AI field. Who argue that although LLMs, AI Tools, and other commercial enterprise creations have upset and opened new economies of scales of talent and resource investing in tech mode growth. They are not necessarily better at ecosystem modeling and simulating dynamics than traditional differentiated techniques employing computation in the loop or as tool. This section offers more detail and nuance on concrete points to consider as we all reflect and direct our future collaborations in research on AI towards socially justified economies of accountability and scale.

5.1. Innovative Approaches to AI Development

The claim from detractors on the bumpy new road opened up to discovery and innovation specific to domain is, that despite how much funding and interest have piggybacked off of the large data model capabilities gap, it does not speak to better capabilities at solving more difficult to solve problems in a concrete domain and context. At developing new innovative differentiated toolsets or protocols specific to those areas or towards associated modeling tasks. The justification and ground

returning observation is similar to that critiquing how thick a philosophical layer open world models and techniques developed in a lab, and the data compare did not result in new easier forms of solving more difficult novel tasks confronting global societies. For some of the same reasons promoted by our Critique of CG, Cycles of Deflation, Technological Innovation and Market Entry. It has been produced and published, endlessly updated and extended, by those it concerns, as becoming acutely aware of and facing these stars of our wars. As collaborative as claim driven resource and investment trajectories have been.

5.2. Interdisciplinary Collaborations

A broader observation and upcoming critique we offer is that in general we feel that outside tech corporations and academia connections and expertise trade are poorly mapped. Partnering in the formation of new communities of talent and aspiring project incubators around shared population learning oriented building cycles. This builds off what is implicit in much of what is said here to further unpack the inter and transdisciplinary conversation feedback loop around model changes that we think is critical to listen to and embrace with regard to both our social fabric and the ecosystem at large.

5.1. Innovative Approaches to AI Development

While AI development has focused on advancing capabilities of general models, future approaches that think more holistically about AI tools and their impact will be needed. In a similar spirit, we argue that AI tools work best when they can both reason about the real world and learn from their interactions with it to adapt behaviors to new domains and contexts without requiring exhaustive external supervision through the entirety of their usage.

To accomplish such a synergistic approach to AI, researchers need to think more creatively about paradigms such as few-shot learning, interactive learning from various types of humans or even other AI actors, AI-assisted semi-supervised or self-supervised learning systems, tools that address concepts like domain transfer and meta-learning, machine learning systems that are adaptable on the fly in real world user-centered contexts, and others. A collaborative, interdisciplinary approach that includes other fields such as robot learning, perceptual control, multi-agent reinforcement learning, AI ethics, among others will help here. Generalizable and adaptive AI are core topics that can be studied and understood using concepts developed in more specialized areas and that have potentially huge impacts in the world.

5.2. Interdisciplinary Collaborations

Over the course of this essay, we have primarily drawn upon ideas and concepts proposed in AI and cognitive science; however, there are insights and tools that other disciplines are advancing that are also relevant. Throughout our lives, we have vicariously borrowed from philosophy, linguistics, sociology, neuroscience, anthropology, psychology, enactivism, and economics, among others. Going forward, we feel there is an immense value to be gained through a greater breadth of direct collaborations. Interdisciplinary collaborations are advantageous to all disciplines involved, as we are all exploring what intelligence is, how it can be developed — whether in natural or artificial systems — and how it can safely support and empower humanity.

While this may seem idealistic, recent trends in linguistics, sociology, and psychology argue that the increasing degrees of automation are contributing to the detriment of aspects of our intelligence. Research actually shows there is far less disagreement within and across disciplines than one may expect, and that the areas of disagreement are being actively explored. Such opportunities for collaborative growth are all around us. However, interdisciplinary collaborations tend to be difficult to sustain, because of slow disciplinary-maturity and the depth of expertise required to produce interesting new research directions. This maturity comes from intense specialization, which is ironically often the preferred strength of disciplines, yet caretaking of disciplines frequently comes at the detriment of innovative endeavors. It is really up to each department to decide. Some universities have responded by allowing for exploratory positions that can often be the bridge to further collaborative explorations.

5.3. Policy Implications for AI Governance

Several key implications for AI governance can be drawn from the arguments. The first is that different domains may require vastly different tailored ethical guidelines. Take, for example, the massive progress in the past few years and forthcoming capabilities of different modalities, including creative multimodal and temporal generalization in GANs, simulated task completion in language models, releasing fears about the economic viability of AGI in the mid-2020s, and eventual military interest in autonomous deepfakes – upon these capabilities being achieved, there may not be an effective scoping process at play. The need for well-timed policies that guide and contain the research of these technologies as per risk level is thus urgent. Regulatory models must also showcase domain diversity. Namely, AI that augments transportation, automated video-conferencing, or remote work must have drastically different updates and approval processes than AI that serves healthcare, space exploration, or military sectors. Simultaneously, multidisciplinary knowledge on best

practices guiding these guidelines is still scarce; as states hurry towards subsidizing companies, they do so without accurate calibration, nor have employed a mindset of inter-agency knowledge pooling. Lastly, and perennial to policy design, safety and quality benchmarks for companies are still arbitrary – no specific taxes, grants, or risk discount guides the public economy towards incentivizing safety and superior technology transfer, or towards disincentivizing economic mass. What are currently treated as stringent company bearer realities – high investment, low churn, people-intensive risks of entry – are more often than not ignored because promises of reduced operational costs alleviate these obstacles for companies.

6. Conclusion

In this short essay we have discussed three key components that can enable the generalization, responsibility, and adaptiveness of AI systems: continuous learning, explicit modeling, and principled design philosophy. We described key challenges and directions in the development of system components for these three areas and presented early results from our work to address some of these challenges. Although individually important, we should note that these three areas are deeply related and their cardinality should not be neglected. Because of the nature of AI systems, progress in these areas must balance the usability, efficiency, and capability of these systems, and prioritize performance in the tasks that are important to the user. Finally, we discussed some aims of our research towards making AI systems more useful and usable. We believe that these principles promote useful generalizable AI systems that will augment the human experience and push boundaries in various domains.

We also acknowledge the problems surrounding the responsible and ethical use of AI techniques. While we address challenges directly related to the generalizability, responsibility, and adaptable use of AIs in complex environments, we believe that these are just as important as specific applications and domains. AI systems are only useful if they solve a real problem and we found it useful to bring some concrete examples to discuss in this review. Our work is motivated by the aforementioned inquiries and we hope through interdisciplinary work we can enable useful and principled systems in the future.

References

[1] Chaddad A, Lu Q, Li J, Katib Y, Kateb R, Tanougast C, Bouridane A, Abdulkadir A. Explainable, domain-adaptive, and federated artificial intelligence in medicine. IEEE/CAA Journal of Automatica Sinica. 2023 Mar 28;10(4):859-76.

[2] Chaddad A, Lu Q, Li J, Katib Y, Kateb R, Tanougast C, Bouridane A, Abdulkadir A. Explainable, domain-adaptive, and federated artificial intelligence in medicine. IEEE/CAA Journal of Automatica Sinica. 2023 Mar 28;10(4):859-76.

[3] Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. 2017 Apr;53(4):2786-812.
[4] Giuliani M, Zaniolo M, Castelletti A, Davoli G, Block P. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. Water Resources Research. 2019 Nov;55(11):9133-47.
[5] Rutenberg I, Gwagwa A, Omino M. Use and impact of artificial intelligence on climate change adaptation in Africa. InAfrican handbook of climate change adaptation 2020 Oct 24 (pp. 1-20). Cham: Springer International Publishing.
[6] Tariq MU. Leveraging artificial intelligence for a sustainable and climate-neutral economy in Asia. InStrengthening sustainable digitalization of Asian economy and society 2024 (pp. 1-21). IGI Global Scientific Publishing.
[7] Panda SP. Artificial Intelligence Across Borders: Transforming Industries Through Intelligent Innovation. Deep Science Publishing; 2025 Jun 6.
[8] Panda SP, Muppala M, Koneti SB. The Contribution of AI in Climate Modeling and Sustainable Decision-Making. Available at SSRN 5283619. 2025 Jun 1.
[9] Shivadekar S. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing; 2025 Jun 30.
[10] Ballestar MT, Martín-Llaguno M, Sainz J. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. Psychology & Marketing. 2022 Dec;39(12):2273-83.
[11] Rodriguez-Delgado C, Bergillos RJ. Wave energy assessment under climate change through artificial intelligence. Science of the Total Environment. 2021 Mar 15;760:144039.
[12] Panda SP. Securing 5G Critical Interfaces: A Zero Trust Approach for Next-Generation Network Resilience. In2025 12th International Conference on Information Technology (ICIT) 2025 May 27 (pp. 141-146). IEEE.
[13] Bird LJ, Bodeker GE, Clem KR. Sensitivity of extreme precipitation to climate change inferred using artificial intelligence shows high spatial variability. Communications Earth & Environment. 2023 Dec 12;4(1):469.
[14] Ajagekar A, You F. Quantum computing and quantum artificial intelligence for renewable and sustainable energy: A emerging prospect towards climate neutrality. Renewable and Sustainable Energy Reviews. 2022 Sep 1;165:112493.
[15] Rane J, Chaudhari RA, Rane NL. Data Privacy and Information Security in Deep Learning Applications: Risk Assessment and Patient Safety Protocols for Big Data Analytics. Ethical Considerations and Bias Detection in Artificial Intelligence/Machine Learning Applications. 2025 Jul 10:54.
[16] Li JJ, Bonn MA, Ye BH. Hotel employee's artificial intelligence and robotics awareness and its impact on turnover intention: The moderating roles of perceived organizational support and competitive psychological climate. Tourism management. 2019 Aug 1;73:172-81.
[17] Tzuc OM, Gamboa OR, Rosel RA, Poot MC, Edelman H, Torres MJ, Bassam A. Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic

climate using artificial intelligence and global sensitivity analysis. Journal of Building Engineering. 2021 Jan 1;33:101625.
[18] Mohapatra PS. Artificial Intelligence-Driven Test Case Generation in Software Development. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:38.
[19] Imanian H, Hiedra Cobo J, Payeur P, Shirkhani H, Mohammadian A. A comprehensive study of artificial intelligence applications for soil temperature prediction in ordinary climate conditions and extremely hot events. Sustainability. 2022 Jul 1;14(13):8065.
[20] Tian P, Xu Z, Fan W, Lai H, Liu Y, Yang P, Yang Z. Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production. 2024 Oct 10;475:143649.
[21] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public health and epidemiology informatics: can artificial intelligence help future global challenges? An overview of antimicrobial resistance and impact of climate change in disease epidemiology. Yearbook of medical informatics. 2019 Aug;28(01):224-31.
[22] Mohapatra PS. Artificial Intelligence-Powered Software Testing: Challenges, Ethics, and Future Directions. Intelligent Assurance: Artificial Intelligence-Powered Software Testing in the Modern Development Lifecycle. 2025 Jul 27:163.
[23] Suura SR. Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing; 2025 Apr 13.
[24] Fousiani K, Michelakis G, Minnigh PA, De Jonge KM. Competitive organizational climate and artificial intelligence (AI) acceptance: the moderating role of leaders' power construal. Frontiers in Psychology. 2024 Mar 25;15:1359164.
[25] Da Silva RG, Ribeiro MH, Mariani VC, dos Santos Coelho L. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals. 2020 Oct 1;139:110027.
[26] Lozo O, Onishchenko O. The potential role of the artificial intelligence in combating climate change and natural resources management: political, legal and ethical challenges. J Nat Resour. 2021;4(3):111-31.
[27] Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nature Climate Change. 2022 Jun;12(6):518-27.
[28] Chen L, Chen Z, Zhang Y, Liu Y, Osman AI, Farghali M, Hua J, Al-Fatesh A, Ihara I, Rooney DW, Yap PS. Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters. 2023 Oct;21(5):2525-57.
[29] Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. Ai & Society. 2023 Feb;38(1):283-307.
[30] Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters. 2019 Nov 22;14(12):124007.
[31] Singh S, Goyal MK. Enhancing climate resilience in businesses: the role of artificial intelligence. Journal of Cleaner Production. 2023 Sep 15;418:138228.

[32] Kadow C, Hall DM, Ulbrich U. Artificial intelligence reconstructs missing climate information. Nature Geoscience. 2020 Jun;13(6):408-13.
[33] Nordgren A. Artificial intelligence and climate change: ethical issues. Journal of Information, Communication and Ethics in Society. 2023 Jan 31;21(1):1-5.
[34] Leal Filho W, Wall T, Mucova SA, Nagy GJ, Balogun AL, Luetz JM, Ng AW, Kovaleva M, Azam FM, Alves F, Guevara Z. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change. 2022 Jul 1;180:121662.
[35] Luccioni A, Schmidt V, Vardanyan V, Bengio Y. Using artificial intelligence to visualize the impacts of climate change. IEEE Computer Graphics and Applications. 2021 Jan 14;41(1):8-14.
[36] Verendel V. Tracking artificial intelligence in climate inventions with patent data. Nature Climate Change. 2023 Jan;13(1):40-7.
[37] Amiri Z, Heidari A, Navimipour NJ. Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. Energy. 2024 Nov 1;308:132827.
[38] Khan MH, Wang S, Wang J, Ahmar S, Saeed S, Khan SU, Xu X, Chen H, Bhat JA, Feng X. Applications of artificial intelligence in climate-resilient smart-crop breeding. International Journal of Molecular Sciences. 2022 Sep 22;23(19):11156.
[39] Panda SP. Relational, NoSQL, and Artificial Intelligence-Integrated Database Architectures: Foundations, Cloud Platforms, and Regulatory-Compliant Systems. Deep Science Publishing; 2025 Jun 22.
[40] Akomea-Frimpong I, Dzagli JR, Eluerkeh K, Bonsu FB, Opoku-Brafi S, Gyimah S, Asuming NA, Atibila DW, Kukah AS. A systematic review of artificial intelligence in managing climate risks of PPP infrastructure projects. Engineering, Construction and Architectural Management. 2025 Mar 28;32(4):2430-54.
[41] Zhao C, Dong K, Wang K, Nepal R. How does artificial intelligence promote renewable energy development? The role of climate finance. Energy Economics. 2024 May 1;133:107493.
[42] Pimenow S, Pimenowa O, Prus P. Challenges of artificial intelligence development in the context of energy consumption and impact on climate change. Energies. 2024 Nov 27;17(23):5965.