Journal & Proceedings of the Royal Society of New South Wales, vol. 157, part 1, 2024, pp. 109–118. ISSN 0035-9173/24/010109-10

Artificial and human intelligence for scientific discovery

Sally Cripps

Director of Technology, Human Technology Institute and Professor of Mathematics and Statistics, University of Technology Sydney

sally.cripps@uts.edu.au

Abstract

This paper will discuss how we might develop AI systems which, together with our human brain, could transform scientific discovery. In order to do this, we need a definition of AI. AI is defined to be that field or industry which is at the intersection of data, algorithms, embedded in an application for the purpose of assisting decision-making.

What is AI?

The problem with Artificial Intelligence (AI) is its name. It either conjures up pictures of futuristic worlds with killer robots empowered by human intelligence, or is put forward as the solution to all the planet's problems. Neither claim is true, and both are unhelpful, (Brooks, 2023). These extreme views are fueled by the media. Reuters on May 30th this year ran the headline:

> Top artificial intelligence executives including OpenAI CEO Sam Altman on Tuesday joined experts and professors in raising the "risk of extinction from AI," which they urged policymakers to equate at par with risks posed by pandemics and nuclear war.

Needless to say these "top artificial intelligence executives" are not a random sample of AI experts. On the contrary, they are a very biased subset, selected precisely because they hold a particular point of view: one that makes headlines[1] But the fact that AI is over-hyped does not mean that it is not useful, nor does it mean that we should be complacent about its misuse.

This paper will discuss how we might develop AI systems which, together with our human brain, could transform scientific discovery. In order to do this, we need a definition of AI. For the purpose of this paper, AI is defined to be that field or industry which is at the intersection of data, algorithms, embedded in an application for the purpose of assisting decision-making. It will also be helpful to categorise AI techniques into two categories. The first category consists of those techniques for which the primary purpose is to make accurate predictions. These techniques will be referred to as predictive AI. They are primarily data-driven, based on neural network architecture, and do not distinguish between cause and effect. The second category consists of those techniques whose primary purpose is to untangle cause and effect, by either encoding a model about the world, or by embedding experiments within the algorithm to infer causation. These techniques will be referred to as causal AI. We note that the two categories are not mutually exclusive: causal AI techniques also give predictions and predictive AI

1 Expert Survey on Progress in AI found that only 5% of AI experts in 2022 (defined to be authors who publish in Neurips or ICML) surveyed stated that AI presented an existential risk.

techniques often attempt to infer causation. Both categories play important and complementary roles in scientific discovery.

Predictive AI

One of the most advanced types of predictive AI is ChatGPT. ChatGPT belongs to a class of algorithms known as Large Language Models (LLMs). It uses data in the form of written text to select the next word in a sentence. When asked to define itself, ChatGPT came back with the following:

> ChatGPT, is an example of Narrow AI, also known as Weak AI or Artificial Narrow Intelligence (ANI). It is designed for specific natural language processing tasks, such as generating human-like text responses, answering questions, and engaging in text-based conversations. ChatGPT, while highly advanced and capable of generating coherent and contextually relevant text, is limited in that it lacks a true understanding of the text it generates.

ChatGPT's acknowledgement that "it lacks a true understanding of the text it generates" is insightful. As an example of this lack of understanding consider the following example from Marcus (2022),

> If you ask LLMs to explain "why crushed porcelain is good in breast milk," they may tell you that "porcelain can help to balance the nutritional content of the milk, providing the infant with the nutrients they need to help grow and develop."

ChatGPT's response sounds authoritative and plausible, but is incorrect. The issue is that the objective function of LLMs is fluency not accuracy. ChatGPT states that its fluency is developed by "relying on patterns and information learned from a massive amount of text data during its training." This is done by a Deep Learning (DL) system that computes associations between words, in the context of a phrase or sentence. LLMs are brilliant at predicting within-sample or interpolating. The success of LLMs in doing this demonstrates that despite the complexity of language, given enough useful data, LLMs can predict what word goes next in sentence, and to construct entire paragraphs which are fluent and plausible text.

Other types of predictive AI include image processing techniques, such as Convolutional Neural Networks (CNNs) (Lecun et al., 1998). Again, impressive as these algorithms are, they make mistakes that a human would never make. Most people in AI and machine learning (ML) have seen a picture similar to that of Figure 1, where adding a small amount of noise to an image can fool the classifier that a pig is now an airliner. Again the reason that these techniques make such mistakes is that, unlike humans, they have no model of the world built into them and have no ability for abstraction and so rely entirely on the information on which they were trained.

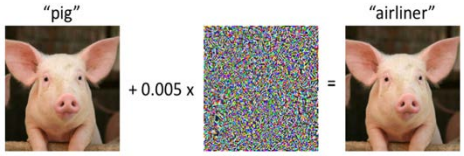

Figure 1: A predictive AI technique correctly classifies the left-hand picture as pig but with a small amount of (non-random) noise added, the same technique now classifies the right-hand picture as an airliner.

However the remarkable achievements made in predictive AI are certainly useful in scientific discovery: their ability to

predict text, based on consuming a large corpus, can be used to summarise existing knowledge, a first step in the process of scientific discovery. The predictive ability of image processing techniques such as CNNs, Generative Adversarial Networks (GAN), (Goodfellow et al., 2014) and Variational AutoEncoders (VAE), (Kipf and Welling, 2016), together with advancements in sensor technology and robotics enables us to capture and analyse data in locations that were previously inaccessible to humans. This is enormously important for scientific discovery.

Figure 2: The Greek characters *πορφυρασ* spell the word *porphyras*, meaning purple in ancient Greek. The Vesuvius Challenge.

Recently it was announced that a machine learning algorithm had deciphered the word "purple" on a Roman scroll from the city of Herculaneum, see Figure 2, carbonised following the eruption of Mt Vesusius. 79 C.E.(*The Economist*, 2023).[2] Yet, although the machine learning algorithm was able to correctly classify the word as "purple," it has no understanding of ancient Greek or English.[3]

Generating accurate predictions does not necessarily lead to generating knowledge or insight. To give another example a Deep Learning system may predict the movement of stars without discovering the underlying laws of nature e.g. gravity, that determine those movements. If AI is to revolutionise scientific discovery it needs to overcome these shortcomings: Predictive AI models, impressive as they are, are not game changers in scientific discovery. They do not incorporate a model of the world, and their treatment of uncertainty is rudimentary at best but most commonly non-existent.

Towards embedding known models of the world

The development of AI techniques that incorporate our knowledge or belief of the world and therefore may be useful in causal inference and scientific discovery is already underway. Physics Informed Neural Networks PINNs (Raissi et al., 2019), are an example. PINNs incorporate models of the world by defining loss functions which penalise solutions which deviate from the physical model. Figure 3, modified from (Karniadakis et al., 2021), is a graphic representation of a PINN for the viscous Burgen system of equations, used in fields such as fluid dynamics. In Figure 3, x represents spatial co-ordinates, t is time, u and $\hat{u}$ are the measured and predicted speeds of the fluid at location x and time t, and v is the viscosity of the fluid. The usual mean squared error (MSE) loss function used to train neural networks, $\mathcal{L}_{NN}$ has been replaced by a weighted average of $\mathcal{L}_{NN}$ and a loss func-

2 First word discovered in unopened herculaneum scroll. https://scrollprize.org/firstletters. Accessed: 2023-10-31.

3 In early 2024: the Vesuvius Challenge 2023 Grand Prize was awarded: we can read the first scroll! https://scrollprize.org/grandprize [Ed.]

tion which penalises solutions which are far from the physics, $\mathcal{L}_{PDE}$, where the partial derivatives required to compute $\mathcal{L}_{PDE}$ are calculated using automatic differentiation techniques. By combining both information from physics and data, these types of models have the potential to shed more light on an issue than either source of information alone. PINNs have been applied to a diverse range of fields including including energy (Hu and You, 2023) and ecology (Robinson et al., 2022).

While this is an exciting area of research, two points should be noted. The first is that the surrogate model $\hat{u}(x,t)$ does not impose the constraints which arise from the physical system, it only penalises solutions which are far from the physics. The second point is that the surrogate model, like many predictive AI techniques which rely on deep learning architecture, are not interpretable, and insights into the scientific phenomenon are limited.

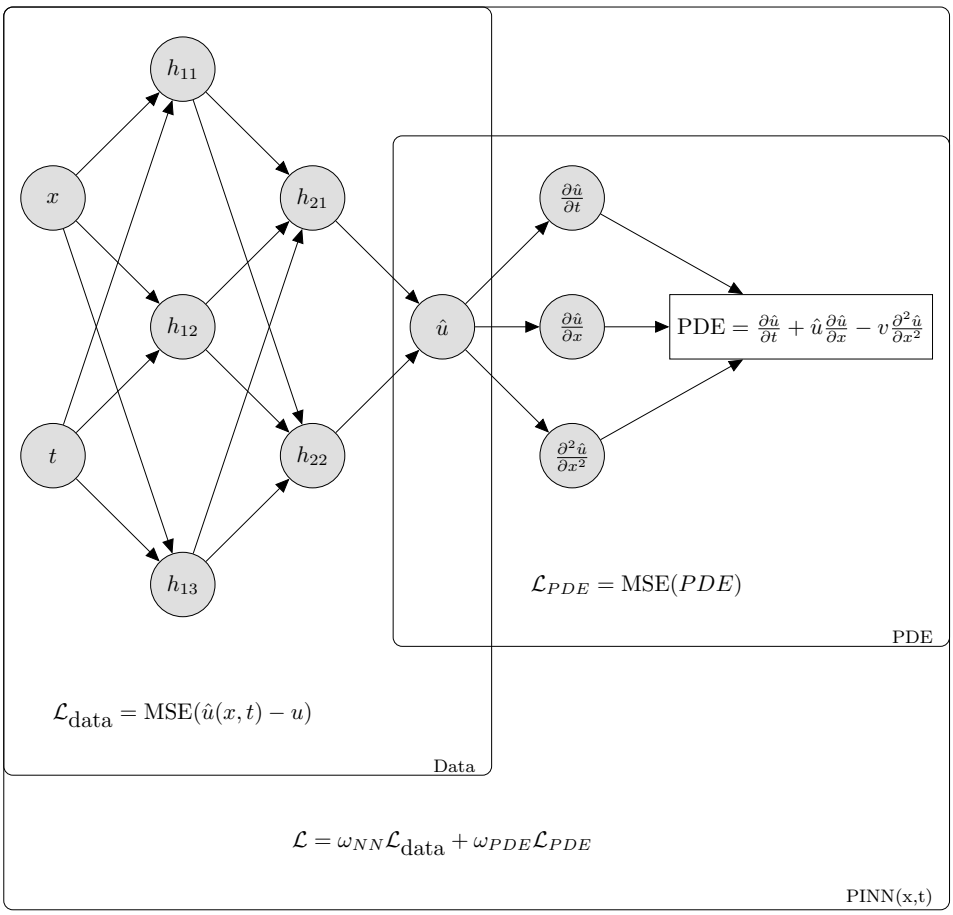

Figure 3: Physics Informed Neural Network (PINN). Graphical representation of estimating the velocity of a fluid u as a function of space x and time t (left box) and the constraints given by the physics of the system (right box). The loss function, $\mathcal{L}$, is a weighted combination of the loss functions of the fit of the neural network (NN) to the data $\mathcal{L}_{data}$ and the fit of the NN to the PDE, $\mathcal{L}_{PDE}$.

Another methodology for incorporating world views into machine-learning techniques is the Bayesian methodology. Indeed the neural network and physics loss functions of PINNs have elements of the Bayesian framework: the neural network loss function is analogous to a likelihood and the physics loss function is analogous to a prior.

The benefit of the Bayesian framework is that it is logically consistent, provides estimates of uncertainty via the posterior distribution and a formal framework which can be generalised to a large class of problems. The partial and ordinary differential equations (PDEs and ODEs), that define many physical systems, such the viscous Burgen system described above, can be expressed as a directed graph, which we denote generically by $\mathcal{G}$.

An example is given in Figure 4, which shows the Lotka Volterra (LV) equations for coral reef growth as a graphical model. The population of coralgal assemblages x, the growth rates by ε and carbonate production by C. The interaction between assemblages denoted by α. Sediment input (Sed), water flow (flow) and depth(Dep) are the basic environmental factors influencing coral growth, via the function f(environ), and the growth rate ε_i is scaled by this factor, see (Salles et al., 2018) and (Pall et al., 2020).

Assuming the physics of coral reef formation are governed by the LV equations, i.e. assuming we know $\mathcal{G}$, the quantities of interest maybe the growth rates ε, the competition matrix A as well as the function which maps the impact of environmental functions to the coral population, $\mathbf{f}$(environ), and estimates and inference of these quantities, jointly denoted by $\theta_{\mathcal{G}} = (\varepsilon, \mathrm{A}, \mathbf{f})$, is via the posterior distribution $p(\theta_{\mathcal{G}} \mid \text{data}, \mathcal{G})$, conditional on the graph $\mathcal{G}$.

It is important to highlight that embedding the LV equations, or any other physical model, is equivalent to assuming that the relationship between factors in system is given by the directed graph structure, with probability one. In a Bayesian setting we express this knowledge as a *prior* distribution, i.e. $P(\mathcal{G} = 1)$, so that there is no uncertainty about this graph structure. However, much of scientific discovery is about uncovering the causal structure of a phenomenon, not just the parameters, θ, of that causal structure, by placing a prior distribution over $\mathcal{G}$, s.t. $\mathcal{G} \sim Q(.)$, where Q is a distribution.

Learning unknown models of the world

The potential for discovery in science has driven much research to learn the structure of a class of graphs known as Directed Acyclical graphs (DAG) (Kitson et al., 2023). The requirement that the graph is acyclical because we wish to infer causation from observational data, (Pearl, 1995), and cycles in the graph structure would make that impossible. We note that causation is only w.r.t an equivalence class (Verma and Pearl, 1990) and only possible if all relevant factors are included in the graph, a condition that is rarely met, so caution is warranted (Dawid, 2010).

Despite these caveats, learning the structure of a graph can provide insight into phenomenon of interest. Consider for example Figure 5, from (Zhu et al., 2023) which depicts the causal structure for a child's Body Mass Index (BMI), denoted by a red diamond in Figure 5. Figure 5 sheds some light on why inventions which target proximal and intermediate causes of child-

hood obesity, such as activity and food type consumption, have not had the impact that might have been expected. Figure 5 clearly shows that childhood obesity is a by-product of social disadvantage, its root causes are socio-economic status (SE) and parental education levels (PE1, and PE2) and that tackling downstream and intermediate factors such as high fat (HF), high sugar (HSD), fruit and vegetable consumption (FV) and activity (FTA) while ignoring these root causes is not sufficient to address the issue, see (Zhu et al., 2023) for a full discussion.

Learning the structure of a graph is an enormously difficult problem. First, the number of possible graphs grows super-exponentially with the number of factors and the space of all possible graphs is discrete, making it difficult to explore the posterior distribution of the graph.

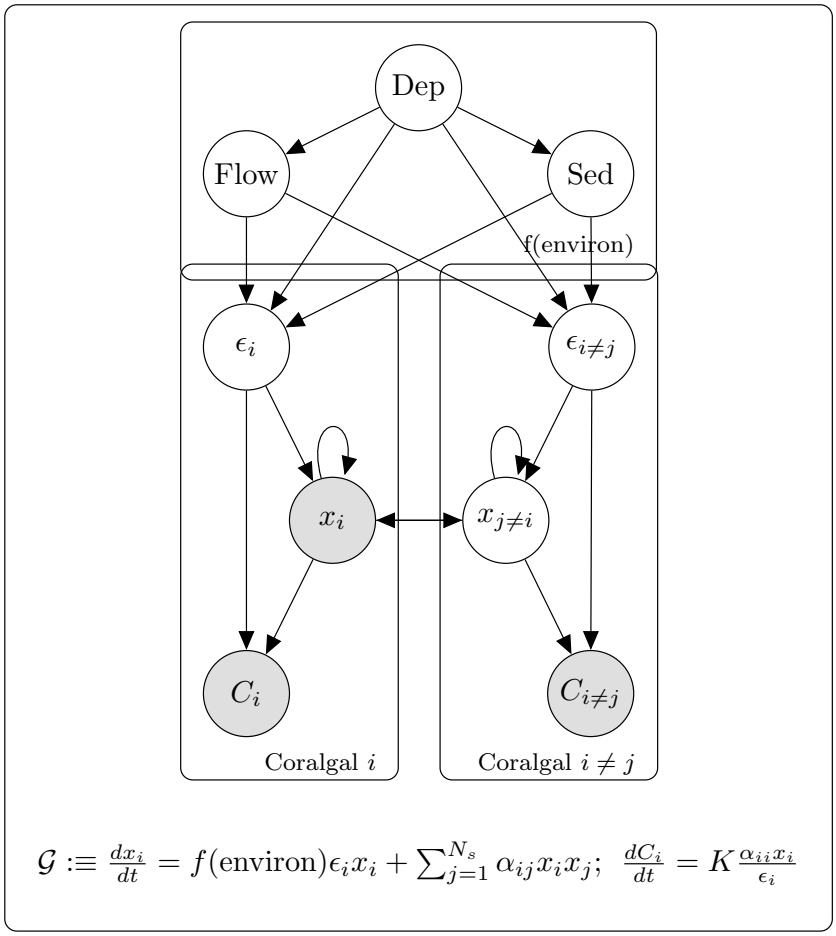

Figure 4: The Lotka Volterra equations depicted as a graphical model. The population of coralgal assemblage i is denoted by x_i, its growth rate by ε_i and its carbonate production by C_i. The interaction between assemblages i and j is denoted by α_{ij}. Sediment input (Sed), water flow (flow) and depth (Dep) are the basic environmental factors influencing coral growth, via the function f(environ), and the growth rate ε_i is scaled by this factor.

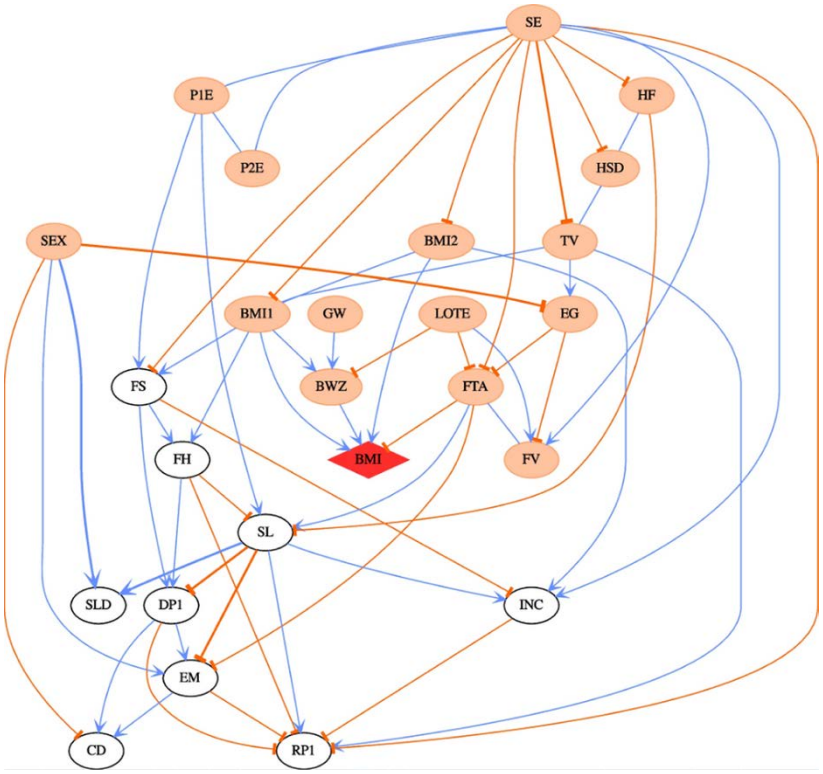

Figure 5: The completed partially directed acyclic graph (CPDAG) derived from the exquivalence class of most probable DAG for 8–10 year olds in the birth cohort of the LSAC (Mohal et al., 2020). The child BMI node is highlighted by a red diamond shape. The thicknesses of the edges in the network correspond to the strength of relationship between nodes, with a thicker line denoting a higher absolute value. The edge coefficients are obtained by regression analysis given the DAG structure. The coefficients of undirected edges are inherited from the values of directed edges. The blue and orange edges indicate positive and negative relationships, respectively. Orange ellipse nodes denote ancestors of child BMI (Zhu et al., 2023).

Second, for Bayesian networks such as in (Zhu et al., 2023), the structure learning algorithms can only learn up to a DAG's equivalence class, in which all the DAGs are equally likely (Verma and Pearl, 1990), and DAGs which belong to the same equivalence class can have very different causal structures. The only way to confirm causality is via experimentation.

Knowledge Discovery Systems: Algorithms which tell us what we don't know

The examples given throughout this paper show how both types of AI techniques, predictive and causal, can be used to aid scientific discovery. The former harness the computational advantages that have developed over the last 50 years, while the latter leverage developments in inferential thinking, developed over the last 100 years, and both are underpinned by developments

in new mathematical methods. However impressive these achievements may be, they are isolated competencies and neither alone will provide a step change in scientific discovery. If we are to transform science with AI, we need to take a systems approach and combine these techniques in a framework capable of knowledge discovery.

What should this system look like? It starts with the development of algorithms which quantify uncertainty. Why? Because uncertainty tells us what we don't know. We need AI systems that, given a question, can assist in the identification and acquisition of valuable information, that can fuse different sources of information in a principled manner, that can produce not only predictions but also levels of confidence to guide real-time experimental design, that can update hypotheses and suggest new ones.

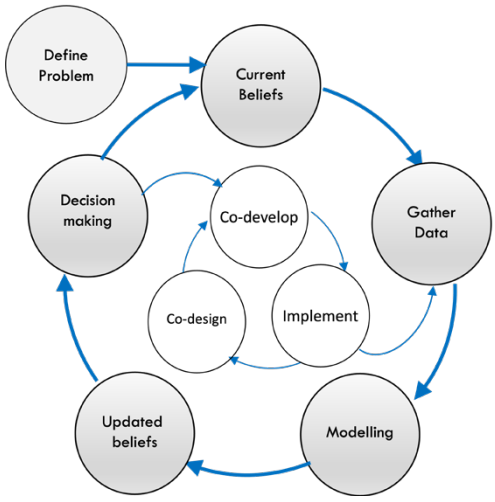

Figure 6: Conceptual AI framework for scientific discovery

Figure 6 is a conceptual AI framework for scientific discovery. It couples ideas from Bayesian reasoning with the growing area of research on collective intelligence and highlights six key features:

1. *Define Problem: scientific discovery starts with specific questions* about unknown quantities, denoted here by $\mathcal{G}$, the causal structure, and the parameters of that structure $\boldsymbol{\theta}_{\mathcal{G}}$. These questions need to be the centre of an AI framework in which evidence gathering, algorithmic and model advancement, system development and decision making are connected in a continuous, iterative, learning cycle.
2. *Current Beliefs: Collecting evidence on what is already known to form* $p(\mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}})$. Predictive AI techniques, such as LLMs can be used to probabilistically summarise what is already known. Additionally, prior elicitation methods (Falconer et al., 2022), which convert varying subjective beliefs from experts or communities into probability distribution, can be combined with more traditional sources of data to gain insights that neither source of information alone could provide.
3. *Gather Data: it is valuable information, not big data that counts.* Scientific discovery proceeds by identifying information gaps and conducting experiments to resolve uncertainties. To accelerate this process, we need algorithms which accurately quantify uncertainty, then, using this estimate of uncertainty, assess the value of future data sources based on their ability to reduce uncertainty concerning the question at hand. One method of doing this is to sequentially acquire data, $\mathcal{D}^*$, that is maximally informative about $\mathcal{G}$, and $\boldsymbol{\theta}_{\mathcal{G}}$, measured for example by the expected mutual information $I(\mathcal{D}, \{\mathrm{G}, \Theta\})$ where

$$I(\mathcal{D}, \{\mathrm{G}, \Theta\}) = \sum_{\mathcal{G} \in \mathrm{G}} \int_{\theta_{\mathcal{G}} \in \Theta} P(\{\mathcal{G}, \theta_{\mathcal{G}}\} \mid \mathcal{D}) P(\mathcal{D}) \log\left(\frac{P(\{\mathcal{G}, \theta_{\mathcal{G}}\} \mid \mathcal{D})}{P(\{\mathcal{G}, \theta_{\mathcal{G}}\})}\right) \mathrm{d}\theta_{\mathcal{G}}$$

and

$$\mathcal{D}^* = \arg\max_{\mathcal{D}\in D} I(\mathcal{D}, \{\mathrm{G}, \Theta\}).$$

New sensor technologies and other data capture techniques, together with predictive AI techniques such as CNNs, VAEs, and GANs, can be used to record and analyse data from a variety of sources.

4. *Modelling: Construct likelihood models*, using multiple sources of data and capable of inferring casual pathways,

$$P(\mathcal{D} \mid \mathcal{G}, \theta_{\mathcal{G}}) = \prod_{i=1}^{N} P(\mathbf{x}_i \mid Pa_i, \theta_{\mathcal{G},\mathcal{G}})$$

where $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$.

5. *Update Beliefs*. The existing information contained in $p(\mathcal{G}, \theta_{\mathcal{G}})$ is combined with information in the new data via the likelihood function $P(\mathcal{D} \mid \mathcal{G}, \theta_{\mathcal{G}})$ in a probabilistic framework, quantifying and updating uncertainty dynamically as new information and discoveries emerge to yield the posterior belief

$$P(\theta_{\mathcal{G}}, \mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) P(\theta_{\mathcal{G}} \mid \mathcal{G}) P(\mathcal{G})}{P(\mathcal{D})}$$

6. *Decision Making*. The incorporation of values, desired outcomes and measure of success is incorporated in the action a from a set of possible actions $\mathcal{A}$, via the formation of the utility function,

$$a^* = \arg\max_{a\in\mathcal{A}} U(a, \mathcal{D}),$$

where $U(a, \mathcal{D})$ is given by

$$U(a \mid \mathcal{D}) = \sum_{\mathcal{G}\in\mathrm{G}} \int_{\theta_{\mathcal{G}}\in\Theta} u(a \mid \mathcal{G}, \theta_{\mathcal{G}}, \mathcal{D}) P(\theta_{\mathcal{G}} \mid \mathcal{G}, \mathcal{D}) P(\mathcal{G} \mid \mathcal{D}) \, \mathrm{d}\theta_{\mathcal{G}}.$$

This is a learning-as-we-go approach: actions are adaptively chosen as new information comes to hand to maximise some prespecified criteria. What is enabled by AI is the identification of valuable information via algorithms that can quantify what we don't know, and the ability to gather and store that information at a rate and in places where it may be difficult for humans to do. In these types of AI systems, the models developed are explainable and transparent. The assumptions are explicit, and therefore the impact of assumptions can be assessed. They are mathematically rigorous and can offer guarantees. They are not just based on associations between factors but are designed to estimate causal pathways so that the right intervention is implemented at the optimal time. And they incorporate human values by being co-designed and co-implemented by the communities which are impacted. In these AI systems, the human is not just in-the-loop, the human is at-the-helm.

References

Brooks R (2023) Just calm down about ChatGPT already,and stop confusing performance with competence. *IEEE Spectrum*. Epub ahead of print 2023.

Dawid AP (2010) Beware of the DAG! In: *Causality: Objectives and Assessment*, pp. 59–86. PMLR.

Falconer JR, Frank E, Polaschek DLL, et al. (2022) Methods for eliciting informative prior distributions: A critical review. Available at: https://arxiv.org/abs/2112.07090

Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Hu G and You F (2023) An AI framework integrating physics-informed neural network with predictive control for energy-efficient food production in the built environment. *Applied Energy*, 348.

Karniadakis GE, Kevrekidis IG, Lu L, et al. (2021) Physics-informed machine learning. *Nature Reviews Physics* 3(6): 422–440.

Kipf TN and Welling M (2016) Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*. Epub ahead of print 2016.

Kitson NK, Constantinou AC, Guo Z, et al. (2023) A survey of bayesian network structure learning. *Artificial Intelligence Review* 56(8): 8721–8814.

Lecun Y, Bottou L, Bengio Y, et al. (1998) Gradient-based learning applied to document recognition. *Proc. of the IEEE* 86(11). Ieee: 2278–2324.

Marcus G (2022) AI platforms like ChatGPT are easy to use but also potentially dangerous. *Scientific American*. Epub ahead of print 2022.

Mohal J, Lansangan C, Gasser C, et al. (2020) *Growing Up in Australia: The Longitudinal Study of Australian Children — Data User Guide. Release 9C1*. Epub ahead of print 2020.

Pall J, Chandra R, Azam D, et al. (2020) Bayesreef: A bayesian inference framework for modelling reef growth in response to environmental change and biological dynamics. *Environmental Modelling & Software* 125: 104610.

Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82(4): 669–688.

Raissi M, Perdikaris P and Karniadakis G (2019) Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 378: 686–670.

Robinson H, Pawar S, Rasheed A, et al. (2022) Physics guided neural networks for modelling of non-linear dynamics. *Neural Networks* 154: 333–345.

Salles T, Pall J, Webster JM, et al. (2018) Exploring coral reef responses to millennial-scale climatic forcings: Insights from the 1-d numerical tool pyReef-core v1.0. *Geoscientific Model Development* 11(6): 2093–2110.

The Economist (2023) AI could help unearth a trove of lost classical texts. Epub ahead of print 2023.

Verma T and Pearl J (1990) Equivalence and synthesis of causal models. In: *Proceedings of the sixth annual conference on uncertainty in artificial intelligence*, USA, 1990, pp. 255–270. UAI '90. Elsevier Science Inc.

Zhu W, Marchant R, Morris RW, et al. (2023) Bayesian network modelling to identify on-ramps to childhood obesity. *BMC Medicine* 21(1): 105.