ADVISORY
COUNCIL

La Playa Beach & Golf Resort
Naples, FL
January 22–24, 2025

Gen Re®

A Berkshire Hathaway Company

Proprietary Notice

The material contained in this presentation has been prepared solely for informational purposes by Gen Re. The material is based on sources believed to be reliable and/or from proprietary data developed by Gen Re. This information does not constitute legal advice and cannot serve as a substitute for such advice. The content of the presentation is copyrighted. Reproduction or transmission is only permitted with the prior consent of Gen Re.

Gen Re®

ADVISORY
COUNCIL

Insurance Decision-Making in the Age of Generative AI

Frank Schmid

Chief Technology Officer

Note: The cartoon of Sir Isaac Newton was hand-drawn by Gen Re's AI engineer Amrita Anam, PhD.

Outline

- A Brief History of Generative AI
- Generative AI is General-Purpose Technology
- Implication for Insurance Decision-Making
- Addendum: Machine Learning vs. Physical Models in Weather Forecasting and Climate Simulation

A Brief History of Generative AI

Deep learning — the engine that powers the training of generative AI systems

- The deep learning revolution started with AlexNet winning the 2012 ImageNet Challenge.
  - ImageNet was an annual contest where research teams around the world had their machine learning algorithms compete in classifying a vast body of images.[1]
  - In the 2012 ImageNet Challenge, AlexNet, developed at the University of Toronto, was the only neural network among the seven contestants.[2,3]
  - AlexNet won the challenge by a wide margin—besides, it was the fastest learning model by far.
- In traditional machine learning, for an algorithm to perform feature recognition in a classification task, it must be instructed what to pay attention to in a process known as feature engineering.
  - Deep learning, on the other hand, learns of the presence of features, a process known as feature learning.
  - For instance, a deep neural network learns on its own the defining features of a cat from labeled cat images (feature learning) and then recognizes these features (feature recognition).

1) The ImageNet Challenge was pioneered by Li Fei-Fei of Stanford University. See https://image-net.org/about.php.

2) The other contestants were teams from the University of Tokyo, Oxford University, University of Jena, University of Amsterdam, and two teams from Xerox Research Europe. See ImageNet, https://image-net.org/challenges/LSVRC/2012/results.html.

3) The University of Toronto team was led by Geoffrey Hinton and counted as team members his students Ilya Sutskever (later co-founder of OpenAI) and Alex Krizhevsky.

A Brief History of Generative AI

The role of NVIDIA Graphics Processing Units (GPUs)

- AlexNet in 2012 was one of the first classification models to run on NVIDIA GPUs.
  - AlexNet was trained on two NVIDIA circuit boards. In comparison, a neural network that Google had trained earlier that year to identify videos of cats required some 16,000 Central Processing Units (CPUs).[1]
  - Developed for graphics processing in video gaming, a GPU breaks down complex mathematical tasks into many small calculations and then processes them in parallel, all at once.
  - The repurposing of NVIDIA GPUs for generalized computing was made possible by NVIDIA's proprietary parallel computing software layer CUDA, released in 2006.
- CUDA was created by Ian Buck, who in 2003 presented a forerunner at a workshop in San Diego.[2]
  - Following the completion of his PhD at Stanford University in 2004, Ian Buck joined NVIDIA.[3]
  - Says Ian Buck, in around 2012 "AI found us."[4]

CUDA: Compute Unified Device Architecture.

[1] *See Stephen Witt, "How Jensen Huang's Nvidia Is Powering the A.I. Revolution," The New Yorker, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution.*

[2] *See Ian Buck, "Data Parallel Computing on Graphis Hardware," Graphics Hardware 2003, San Diego, https://graphics.stanford.edu/~ianbuck/GH03_datapargfx.pdf. The 2003 events that led to the repurposing of GPUs for generalized computing were captured by Michael Macedonia (2003) "The GPU Enters Computing's Mainstream," Georgia Tech Research Institute, https://users.cs.northwestern.edu/~ago820/cs395/Papers/W05_HardwareRendering/GPUMainStream.pdf.*

[3] *An early sponsor of Ian Buck's work at Stanford University was DARPA (Defense Advanced Research Project Agency). See Stanford University,* http://graphics.stanford.edu/~ianbuck/.

[4] *See Stephen Witt, op. cit.*

A Brief History of Generative AI

The Transformer

- The Transformer was introduced by Google Brain in 2017.[1]
- Now the dominant neural network architecture, the Transformer is based on a self-attention mechanism, which directly models relations among all words in a sentence.[2,3]
  - For example, deciding on the most likely meaning of the word "bank" in the sentence "I arrived at the bank after crossing the…" requires knowing if the sentence ends on "street" or "river."
  - To determine that the word "bank" refers to the shore of a river and not a financial institution, the Transformer can learn to attend immediately to the word "river" and make this decision in a single step.
- Unlike older neural network architectures, the Transformer is scalable, horizontally and vertically.
  - The self-attention architecture of the Transformer lends itself to parallelization.
  - The Transformer is a solution to the vanishing gradient problem, allowing the stacking of a high number of hidden layers.

1) In 2023, Alphabet merged Google Brain and DeepMind into a newly formed Google DeepMind. See, for instance, Roula Khalaf, "Google's DeepMind'-Brain merger: tech giant regroups for AI battle," Financial Times, April 28, 2023, https://www.ft.com/content/f4f73815-6fc2-4016-bd97-4bace459e95e.

2) For the original paper on the Transformer, see Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is All You Need," arXiv, submitted June 12, 2017, last revised August 2, 2023, https://arxiv.org/pdf/1706.03762.pdf.

3) For a gentle introduction to the Transformer, see Visual Storytelling Team and Madhumita Murgia, "Generative AI exists because of the transformer," Financial Times, September 12, 2023, https://ig.ft.com/generative-ai.

Generative AI is General-Purpose Technology

Three defining characteristics of general-purpose technology (GPT)

- A GPT (1) is widely used, (2) is capable of ongoing technical improvement, and (3) enables innovation in application sectors.[1]
- The arrival of a GPT is a rare event, even in modern times.
  - Examples of older GPTs are the steam engine, the electric motor, and the semiconductor.
- The adoption of a GPT is gradual, and its productivity benefits take time to materialize.
  - The GPT complements innovation in production processes, organizational design, and products.
  - The inventions of the electric motor (ca. 1890) and the personal computer (1981) gave rise to productivity booms in the United States with time lags of 25 and 15 years, respectively.[2,3]

1) *See Timothy F. Bresnahan, "General Purpose Technologies," in: Bronwyn H. Hall and Nathan Rosenberg (eds.) Handbook of the Economics of Innovation 2: 761-791, 2010, https://www.sciencedirect.com/science/article/pii/S0169721810020022.*

2) *See Martin Wolf, "The threat and promise of artificial intelligence," Financial Times, May 9, 2023, https://www.ft.com/content/41fd34b2-89ee-4b21-ac0a-9b15560ef37c. See also Dominic Wilson and Vickie Chang, "Markets around past productivity booms," Top of Mind 120: 18-19, July 5, 2023, https://www.goldmansachs.com/intelligence/pages/top-of-mind/generative-ai-hype-or-truly-transformative/report.pdf.*

3) *For two studies on the possible productivity-enhancing effect of generative AI see Joseph Briggs and Devesh Kodnani, "The Potentially Large Effect of Artificial Intelligence on Economic Growth," Global Economics Analyst, Goldman Sachs, March 26, 2023, https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html; and Michael Chui et al., "The economic potential of generative AI: The next productivity frontier," McKinsey & Company, June 2023, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.*

Generative AI is General-Purpose Technology

The increasing velocity of the feedback cycle of complementary innovation

- A GPT unleashes a feedback cycle of continuing technical improvement and accompanying downstream innovation.[1]
- For generative AI, the velocity of the feedback cycle is expected to be higher than it was for the steam engine, electricity, or electronic computing.
- A seminal study on the transition of U.S. corporations from mainframe computing to C/S (client/server) computing observed that the organizations slowest to transition were those with the highest *cost* of adoption rather than with the lowest *benefit* of adoption.[2]
- Technology choices that allow the insurer to benefit from a high velocity of complementary innovation emphasize the importance of learning and reversibility.[3]

1) See Timothy F. Bresnahan and Manuel Trajtenberg, “General Purpose Technologies ‘Engines of Growth’?” Journal of Econometrics 65 (1): 83-108, 1995, https://www.sciencedirect.com/science/article/pii/030440769401598T.

2) See Timothy Bresnahan and Shane Greenstein, “Technical Progress and Co-Invention in Computing and in the Uses of Computers“ Brookings Papers on Economic Activity: Microeconomics 1996: 1-83, 1996, https://www.brookings.edu/articles/technical-progress-and-co-invention-in-computing-and-in-the-uses-of-computers.

3) The concept of real options valuation delivers the theoretical foundation for principles of technology decision-making that account for the simultaneous presence of learning and irreversibility.

Generative AI is General-Purpose Technology

The potential of a J-curve effect in (measured) productivity

- The arrival of a GPT requires an organization to make complementary investments.
- At the early stage of a GPT adoption process, the organization builds intangible assets.
  - These assets are output that the organization retains (unlike its products and services, which it sells).
  - Once built, these intangible assets serve as input to the organization's production process.
- Intangible assets tend to escape traditional measures of productivity.
  - Productivity may briefly be underestimated before being mildly overestimated for some time.[1]
- A crude measure of the insurer's productivity is the expense ratio.[2]
  - When built, intangible assets are retained output, adding to the numerator. Then, when serving as input in production, they contribute to the denominator.

[1] *See Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," American Economic Journal: Macroeconomics 13(1): 333-372, 2021, Working Paper (January 2020) at https://www.nber.org/papers/w25148.*

[2] *Although the expense ratio can serve as a meaningful measure of productivity for any given insurer, differences in expense ratios across insurers are not necessarily indicative of differences in productivity. This is due to differences in business models.*

Generative AI is General-Purpose Technology

The benefits of augmentation of labor

- The arrival of a general-purpose technology offers opportunity for automation.
  - Automation can augment labor by enhancing and complementing the skills of humans…
  - …and automation can substitute labor.[1]
- Augmentation raises the value of human labor.[2]
  - Substitution and augmentation both occur, but the latter offers the greater economic benefit by far.
  - The market value of an hour of human labor, as measured by median wages, has grown more than tenfold since 1820.
- New types of work arrive[3]
  - About 60 percent of jobs in the United States represent types of work that have been created since 1940.
  - Many new occupations are directly related to new technology—an example is the computer programmer.
  - Some occupations (in health care, leisure, etc.) emerge in response to consumer demand in the wake of rising incomes.

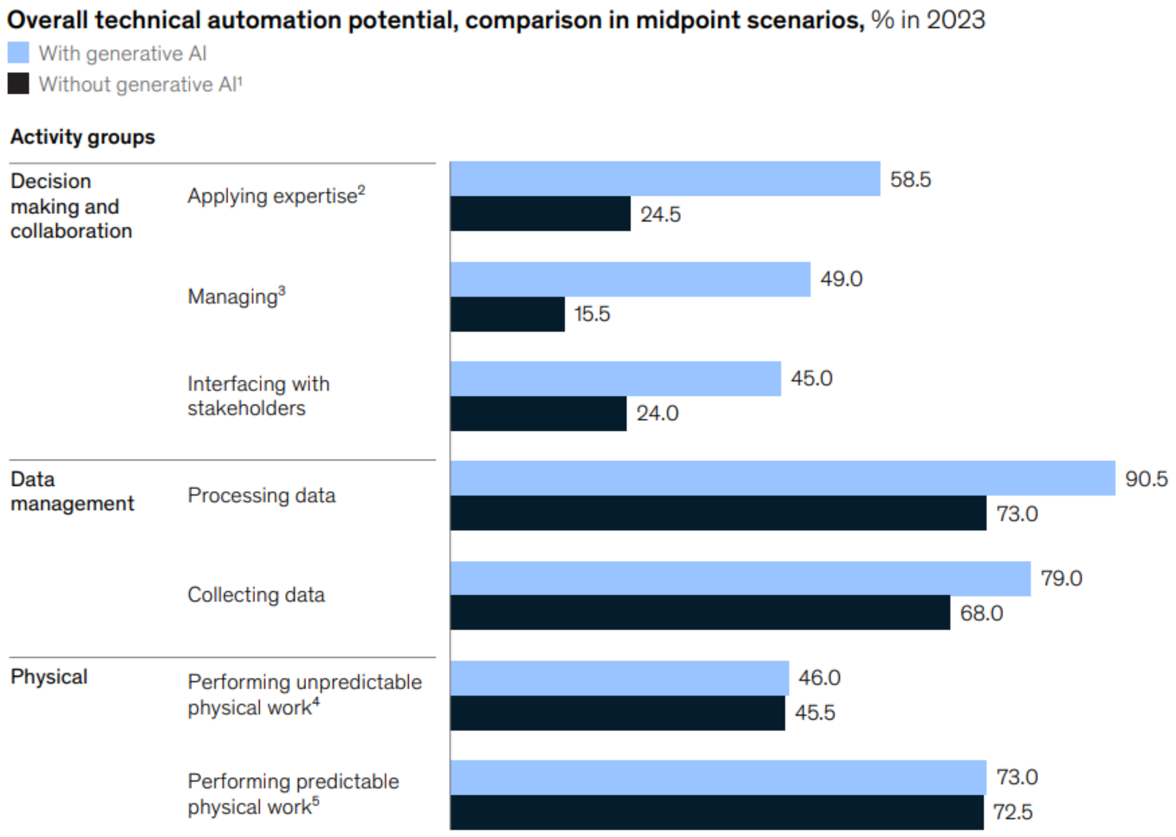
[1] *See David H. Autor, "Why Are There Still So Many Jobs: The History and Future of Workplace Automation," Journal of Economic Perspectives 29(3): 3-30, 2015, https://www.aeaweb.org/articles?id=10.1257/jep.29.3.3.*

[2] *See Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," Daedalus 151(2): 272–287, 2022, https://direct.mit.edu/daed/article/151/2/272/110622/The-Turing-Trap-The-Promise-amp-Peril-of-Human.*

[3] *See David Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller, "New Frontiers: The Origins and Content of New Work, 1940–2018," Quarterly Journal of Economics 139(3): 1399–1465, 2024, https://doi.org/10.1093/qje/qjae008.*

Implication for Insurance Decision-Making

Decision-making and collaboration are the areas where generative AI is expected to have the greatest impact



Note: Figures may not sum, because of rounding.
[1]Previous assessment of work automation before the rise of generative AI.
[2]Applying expertise to decision making, planning, and creative tasks.
[3]Managing and developing people.
[4]Performing physical activities and operating machinery in unpredictable environments.
[5]Performing physical activities and operating machinery in predictable environments.
Source: McKinsey Global Institute analysis

- "Without generative AI" projections (dark blue) refer to the 2017 McKinsey study *Jobs lost, jobs gained: Workforce Transitions in a time of Automation*, available at mckinsey.com.
- It has been posited that generative AI may reduce the scarcity of labor in the category *Decision-making and collaboration: Applying expertise* by enhancing and complementing the decision-making competencies of medium-skilled expert labor.[1]

Source of chart: McKinsey & Company, "The Economic Potential of Generative AI: The Next Productivity Frontier," June 2023, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

1) *See Delphine Strauss, "David Autor: 'We have a real design choice about how we deploy AI'," Financial Times, August 10, 2023, https://www.ft.com/content/9c087da3-63d2-4d73-97dc-023025b529aa.*

Implication for Insurance Decision-Making

"The week that artificial intelligence swept the Nobel Prizes," *Financial Times*, October 12, 2024

- "They used physics to find patterns in information."[1]
  - "This year's laureates used tools from physics to construct methods that helped lay the foundation for today's powerful machine learning. John Hopfield created a structure that can store and reconstruct information. Geoffrey Hinton invented a method that can independently discover properties in data[,] and which has become important for the large artificial neural networks now in use."
- "They cracked the code for proteins' amazing structures."[2]
  - "The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools. David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins. Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins' complex structures. These discoveries hold enormous potential."
  - For background, the referenced AI model is AlphaFold2, developed by Google DeepMind, led by Demis Hassabis.[3]

1) *See The Nobel Prize, Prizes 2024,* https://www.nobelprize.org/all-nobel-prizes-2024/.

2) *Ibid.*

3) *Demis Hassabis co-founded DeepMind in 2010, and Google acquired it in 2014. In 2023, Google' parent Alphabet merged Google Brain and DeepMind into a newly formed Google DeepMind. See, for instance, Roula Khalaf, "Google's DeepMind'-Brain merger: tech giant regroups for AI battle," Financial Times, April 28, 2023, https://www.ft.com/content/f4f73815-6fc2-4016-bd97-4bace459e95e.*

Note: Interestingly, Geoffrey Hinton is the great-great grandson of George Boole (1815-1864). Boolean algebra is foundational for computer science.

Implication for Insurance Decision-Making

Curious aspects of the Nobel Prizes awarded for work on artificial intelligence

- Artificial intelligence is an engineering science where one builds an artifact and then studies it.[1]
  - Artificial intelligence is a tool for scientific discovery, rather than the discovery of a natural phenomenon.
  - Neural networks do not have an overarching theoretical framework although there are important contributing theoretical concepts, such as statistical learning theory, information theory, among others.
- Neural networks make predictions solely based on complex relationships in large datasets.[2]
  - In 1972, Christian Anfinsen was awarded the Nobel Prize in Chemistry for his hypothesis that a protein's structure is dictated by the sequence of its amino acids. This concept became known as Anfinsen's dogma.
  - During the 1960s, Cyrus Levinthal demonstrated that a protein chain could theoretically adopt an enormous number of possible conformations. If a protein were to explore all these conformations, it would take an unimaginably long time, comparable to the age of the Universe. This observation is referred to as Levinthal's paradox.
  - Anfinsen's dogma suggests that predicting a protein's folded state does not necessarily require an understanding of the folding process itself, which allowed Google DeepMind's AlphaFold2 to get around Levinthal's paradox.

1) See Madhumita Murgia, "Google DeepMind's Demis Hassabis on his Nobel Prize: 'It feels like a watershed moment for AI'," Financial Times, October 21, 2024, https://www.ft.com/content/72d2c2b1-493b-4520-ae10-41c1a7f3b7e4.

2) See AlphaFold: A practical guide, What is the protein folding problem?, https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/what-is-the-protein-folding-problem.

Implication for Insurance Decision-Making

A brief history of quantification in decision-making

- Descriptive statistics (Adrien-Marie Legendre and Carl Friedrich Gauss)
  - Legendre (1805) and Gauss (1795) used least squares for describing the movements of celestial bodies.
  - $\bar{y} = \hat{a} + \hat{b} \cdot \bar{x}$, where $\bar{y}$ and $\bar{x}$ are arithmetic means of astronomical measurements, and $\hat{a}$ and $\hat{b}$ are estimated parameters.[1]
- Quantifying causal relations (Structural models; Sir Francis Galton)
  - Galton (1870s), studying heredity in peas, used least squares to relate the size of daughter peas to those of mother peas.[2]
  - $\hat{y} = \hat{a} + \hat{b} \cdot x$, where x causes y, the coefficients $\hat{a}$ and $\hat{b}$ are estimated parameters, and $\hat{y}$ is the prediction.
- Quantifying connections (Deep learning)
  - Deep learning can capture very high-dimensional and complex relations, unconstrained by our knowledge of the world.
  - $\hat{y} = f(x)$, where $f(.)$ is a pre-trained deep neural network (e.g., Google DeepMind's GraphCast for weather forecasting), x is input (e.g., the state of the Earth's atmosphere currently and six hours earlier), and $\hat{y}$ is the prediction.[3]

[1] *A regression line established by means of ordinary least squares travels through the center of the data cloud.*

[2] *See, for instance, Jeffrey M. Stanton (2017) "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors," Journal of Statistics Education 9(3), https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537.*

[3] *See Ilan Price et al., "Probabilistic weather forecasting with machine learning." Nature 637: 84–90, 2025, https://doi.org/10.1038/s41586-024-08252-9.*

Implication for Insurance Decision-Making

Artificial intelligence as an abstract layer

- We don't fully understand how an AI system arrives at its predictions.
  - In this way, artificial intelligence is an abstract layer, a familiar concept in science.[1]
  - For instance, we can understand chemistry in its own abstract layer, without having to understand the physics (quantum mechanics) that lies underneath.[2]
- Abstract layers figure prominently in our daily lives.
  - In learning how to drive a motor vehicle, we construct a layer of abstraction that enables us to drive the engineered artifact without an understanding of the physics powering it.
  - We verify the abstract layer based on the predicted behavior of the artifact.[3]
- Predictions without us understanding the "why" calls for new narrative.
  - Structural models support narrative grounded in causality, delivering meaning and interpretation, whereas neural networks offer only verification of the predictive benefit.[4]

Note: The cartoon of Sir Isaac Newton was hand-drawn by Gen Re's AI engineer Amrita Anam, PhD.

[1] *See Madhumita Murgia, "Google DeepMind's Demis Hassabis on his Nobel Prize: 'It feels like a watershed moment for AI'," Financial Times, October 21, 2024, https://www.ft.com/content/72d2c2b1-493b-4520-ae10-41c1a7f3b7e4.*

[2] *Ibid.*

[3] *See Ellen S. O'Connor (1966) "Telling decisions: The role of narrative in organizational decision making," in: Zur Shapira (ed.) Organizational Decision Making, Cambridge University Press, Chapter 14, 304-323.*

[4] *Ibid.*

Addendum: Machine Learning vs. Physical Models in Weather Forecasting and Climate Simulation

Ex post established connections vs. ex ante postulated causality

Ex Post Connections vs. Ex Ante Causality

Medium-range weather forecasting

- The European Centre for Medium-Range Weather Forecasts (ECMWF) operates HRES, the world's most accurate deterministic physics-based model for medium-range weather forecasting.[1]
  - HRES uses supercomputing to crunch equations based on scientific knowledge of atmospheric physics.
- In 2023, Google DeepMind introduced GraphCast, a graph neural network (GNN).[2]
  - Trained on 39 years of ECMWF historical data, GraphCast predicts weather variables globally up to 10 days ahead.
  - GraphCast has proven to outperform HRES forecasts in 90 percent of 1,380 target variables and demonstrates better severe event predictions for tropical cyclones, atmospheric rivers, and extreme temperatures.[3]
- GraphCast produces a 10-day forecast in less than one minute on a single Google TPU.
  - In comparison, generating forecasts using HRES takes about an hour on a high-end performance computer.[4]

HRES: High-Resolution Forecast (ECMWF, https://confluence.ecmwf.int/pages/viewpage.action?pageId=191632109). TPU: Tensor Processing Unit.

[1] See Remi Lam et al. "Learning skillful medium-range global weather forecasting." Science 382: 1416–1421, 2023, https://www.science.org/doi/epdf/10.1126/science.adi2336. ECMWF also operates ENS, the world's most accurate ensemble model. Ensemble models delivers probability distributions of different future weather scenarios.

[2] Ibid.

[3] In December 2024, Google unveiled GenCast, a generative AI model that uses an ensemble method to generate 15-day forecasts. GenCast has proven to outperform the 15-day forecasts of ENS, ECMWF's ensemble forecasting model, on 97.2 per cent of 1,320 target variables. The See Ilan Price et al., "Probabilistic weather forecasting with machine learning." Nature 637: 84–90, 2025, https://doi.org/10.1038/s41586-024-08252-9.

[4] See Remi Lam et al., op. cit.

Ex Post Connections vs. Ex Ante Causality

Long-range weather forecasting and climate modeling

- In 2024, Google Research introduced NeuralGCM, which bolts generative AI onto existing physics-based general circulation models (in an approach known as differential modeling)[1]
  - General circulation models (GCMs) are foundational for weather and climate prediction.
  - For longer-term predictions, integrating physics with deep learning ensures that the machine learning model observes physical constraints and is capable of generalizing (that is, capable of simulating the unprecedented).
  - NeuralGCM was trained on 80 years of ECMWF data to predict weather up to 3 days ahead.
  - Although trained on medium-range data, NeuralGCM can simulate the atmosphere over a duration of multiple decades.
- NeuralGCM has proven to outperform physics-based X-SHiELD[2] in weather variables related to the seasonal cycle and in predicting emergent phenomena (here, tropical cyclones).[2]
- NeuralGCM generates 70,000 simulation days in 24 hours using a single Google TPU.
  - In comparison, within the same time window, X-SHiELD generates only 19 simulation days, requiring 13,824 CPU cores.[3]

ECMWF: European Centre for Medium-Range Weather Forecasts. TPU: Tensor Processing Unit.

1) See Dmitrii Kochkov et al. "Neural general circulation models for weather and climate." Nature 632: 1060–1066, 2024, https://doi.org/10.1038/s41586-024-07744-y.

2) X-SHiELD (eXperimental System for High-resolution on Earth-to-Local Domains) is developed by the Geophysical Fluid Dynamics Laboratory of NOAA (National Oceanic and Atmospheric Administration), https://www.gfdl.noaa.gov/shield.

3) See Dmitrii Kochkov et al., op. cit.

ADVISORY
COUNCIL

Thank you!

Frank Schmid
frank.schmid@genre.com
203 461 1944