Machine Learning meets Scientific Understanding

Date: 26.06.25, 27.06.25
Place: Internationales Begegnungszentrum, TU Dortmund

Main Organizers: Annika Schuster, Frauke Stoll & Florian J. Boge
Contact: udnn.fk14@tu-dortmund.de

IfPP tu dortmund university UDNN

Emmy Noether-Programm
DFG Deutsche Forschungsgemeinschaft

This workshop is an event organized by the Emmy Noether Group

UDNN: Scientific Understanding and Deep Neural Networks

`https://udnn.tu-dortmund.de/`

Organizers: Annika N. Schuster, Frauke Stoll, Leon Augustin, Levin Burghardt & Florian J. Boge

Contents

About **4**
Machine Learning meets Scientific Understanding 4

Timetable **5**
Thursday, June 26th 5
Friday, June 27th 6

Abstracts **7**
Thursday, June 26th 7
Friday, June 27th 11

Useful Information **19**
General 19
How to get around? 19

Organization and Funding **20**

About

Machine Learning meets Scientific Understanding

The Emmy Noether group UDNN: Scientific Understanding and Deep Neural Networks is pleased to host its second workshop, *Machine Learning Meets Scientific Understanding*, which takes place on June 26th and 27th at TU Dortmund University. This interdisciplinary event brings together philosophers of science and Machine Learning (ML), as well as ML practitioners, to explore the intersections between ML and scientific understanding.

Organizers: Annika N. Schuster, Frauke Stoll, Leon Augustin, Levin Burghardt & Florian J. Boge

Timetable

Thursday, June 26th

09:00	**Arrival**		
10:00		**Introduction**	
10:15–11:00		**Holger Lyre** Magdeburg University	*Semantic Grounding in Advanced LLMs?*
11:15–12:00		**Daniel Neider** TU Dortmund	*A Gentle Introduction to Neural Network Verification (and How It Might Contribute to Evaluating Scientific Insights)*
12:00–13:00	**Lunch**		
13:00–13:45		**Heather Champion** Tübingen University	*On Scientific Discovery With Machine Learning: What is "Strong" Novelty?*
14:00–14:45		**Edward Chang** Stanford University	*From Generative AI to AGI: Multi-LLM Agent Collaboration as a Path Forward*
14:45–15:15	**Coffee Break**		
15:15–16:00		**Emily Sullivan** Utrecht University	*Idealization Failure in ML*
16:15–17:00		**Henk W. de Regt & Eugene Shalugin** Radboud University	*Bridging Scientific Understanding and Creativity with an LLM Benchmark for Narrow-Domain Scientific Fields*

Friday, June 27th

09:00	**Introduction**		
09:15–10:00		**Darrell Rowbottom** Lingnan University Hong Kong	*What's Hidden Inside Predictively Successful Deep Learning Models?*
10:15–11:00		**Sara Pernille Jensen** Oslo University	*The Underdetermination of Representational Content in DNNs*
11:15–12:00		**Timo Freiesleben** Tübingen University	*The Benchmarking Epistemology – What Inferences Can Scientists Draw from Competitive Comparisons of Prediction Models?*
12:00–13:00	**Lunch**		
13:00–13:45		**Giovanni Galli** University of Teramo	*Deep-learning Models and Scientific Understanding through Explanations and Representations*
14:00–14:45		**Insa Lawler** UNC Greensboro	*Machine Learning, AI, and The Gradability of Explanatory Understanding*
14:45–15:00	**Coffee Break**		
15:00–15:45		**Finnur Dellsén** University of Iceland	*Scientific Progress in the Age of AI*
16:00–16:45		**Cameron Buckner** University of Florida	*Predictively-Valid "Alien" Features, or Artifacts? Coping with Inscrutable Scientific Progress*

Abstracts

Thursday, June 26th

Semantic Grounding in Advanced LLMs

Holger Lyre

Magdeburg University

ML models serve as scientific tools and help to deliver our scientific understanding in much the same way as models in general. However, ML models are special in a certain respect: namely, when these models themselves acquire a semantic grounding, when they start to become cognitive and thus possess a form of understanding themselves. Obviously, semantically grounded models will be far more powerful than ungrounded models, they could potentially achieve the status of scientific partners rather than tools. In my talk, I will explore the question of whether advanced LLMs already show signs of semantic grounding, and I will argue that this is indeed the case. To assess the question of semantic grounding, five methodological ways will be distinguished. The most promising way, I claim, is to apply core assumptions of theories of meaning in philosophy of mind and language to LLMs. I will demonstrate that grounding proves to be a gradual affair with a three-dimensional distinction between functional, social and causal grounding. Modern LLMs show basic evidence in all three dimensions. A strong argument is that they develop world models. Hence, LLMs are no stochastic parrots, but already understand the language they generate, at least in an elementary sense.

A Gentle Introductino to Neural Network Verification (and How It Might Contribute to Evaluating Scientific Insights)

Daniel Neider

TU Dortmund

The increasing use of artificial intelligence in safety-critical domains such as autonomous systems and healthcare demands robust methods to ensure the reliability and safety of these technologies. Neural network verification has emerged as a vital research area, providing algorithmic frameworks to rigorously analyze and guarantee critical properties of neural networks across diverse applications. This talk provides an accessible introduction to this critical field, with a specific emphasis on safety-critical properties of neural networks and the algorithmic frameworks designed to automatically verify these properties. Furthermore, we briefly discuss how verification might contribute to evaluating scientific insights obtained through machine learning, inviting exploration of its potential role in advancing scientific understanding.

On Scientific Discovery With Machine Learning: What is ”Strong” Novelty?

Heather Champion

Tübingen University

Recent philosophical accounts of machine learning's (ML) impact negatively appraise the significance of its contribution to changing scientific theory, understanding, or concepts. While valuable, these analyses tend to focus narrowly on one kind of strong disruption claim owing to one notion of "strong" novelty, which is not always clarified. Meanwhile, some philosophers assess whether the capacities of ML algorithms are sufficient to cause one of these disruptions (e.g. whether they utilize creative processes). But what exactly constitutes strongly novel outcomes of ML-enabled science? Omitting a multifaceted answer to this question risks overemphasizing the non-disruptiveness of ML. Also, while the analysis of novel outcomes and the means that successfully achieve them are inevitably linked, outcome desiderata play an important role in evaluating human-computer interactions. Therefore, I focus on outcomes enabled by ML, such as predictions, ideas, or virtual artifacts. I first raise three difficulties with Ratti (2020) and Boge's (2022) outcome-based characterizations of strong novelty: (1) Ratti's domain-specific focus is unnecessary, (2) both underappreciate the scientific impact of token predictions, and (3) Boge is ambiguous about the kind of prior information he takes to be disqualifying for "use novelty," while I argue that on several interpretations, use novelty does not helpfully discriminate strong novelty. However, their accounts capture useful intuitions: changes to existing theory, scientific knowledge, or research direction are highly impacting, as are many outcomes achieved without a certain kind of informational bias (elaborated below).

Next, I introduce a new, wide, variety of outcome-based notions of strong novelty from philosophy of creativity, epistemology, and philosophy of science. I illustrate these with cases from various scientific domains, such as economics and astrophysics. First, a creative outcome generates surprise when an idea with low expected utility turns out to be useful. Alternatively, outcomes that reduce uncertainty ("blindness") regarding an idea's utility helpfully steer research. These notions of belief revision assume a state of awareness regarding a proposition, but ML might also generate this awareness, eliminating deep ignorance regarding scientifically useful patterns, evidence, or hypotheses. Zooming out from these notions of local epistemic change, ML might make broader impact by prompting conceptual change. Particularly, if deep learning methods directly learn conceptualizations useful for specific tasks, these might diverge from existing human conceptualizations. Finally, using ML to learn from data with some independence of local theory regarding a target phenomenon has generative power for scientific change. I define local theory as theory that demarks or explains a target phenomenon. This "bottom-up" form of learning constitutes strong novelty for science because it signals an aim to find a new research direction, often by relying on a different set of cognitive tools for analyzing multidimensional data. It also clarifies that local theory is the kind of prior information that diminishes the generative impact of an ML prediction. My taxonomy clarifies desiderata for scientific exploration with ML and complements assessments of what algorithmic processes might achieve them. It also invites reflection on how some forms of novelty might co-occur and what problems this raises for scientific understanding.

References

[1] Boge, Florian J. "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds and Machines* 32 (2022): 43–75.

[2] Ratti, Emanuele. "What Kind of Novelties Can Machine Learning Possibly Generate? The Case

of Genomics." *Studies in History and Philosophy of Science Part A* 83 (2020): 86–96.

From Generative AI to AGI: Multi-LLM Agent Collaboration as a Path Forward

Edward Chang

Stanford University

The rise of large language models (LLMs) has transformed AI—shifting it from passive analysis to generative capabilities, from narrow task-specific tools to general-purpose systems, and from monolithic models to collaborative multi-agent frameworks. While some experts anticipate the emergence of Artificial General Intelligence (AGI) by 2040, critics like LeCun (2023) argue that LLMs cannot lead to AGI, citing their lack of world models, persistent memory, structured reasoning, and planning capabilities. Critics also highlight how LLMs require massive training data yet still fail to match the efficient few-shot learning demonstrated by even young children.

This talk challenges these critiques by positioning LLMs not as complete solutions, but as necessary substrates for AGI emergence, analogous to how unconscious processes enable conscious reasoning in humans. Just as humans aren't born with blank slates but with neural priors that scaffold learning, LLMs provide foundational capabilities for in-context learning and environmental adaptation. By augmenting LLMs with transactional reliability, self-validation mechanisms, Socratic reasoning, and multi-agent architectures, we can address their current limitations. The proposed Multi-LLM Agent Collaboration framework offers a pragmatic, scalable path toward AGI, where intelligence emerges not from a single model but through structured interaction, persistent memory, and collective reasoning across networked systems.

Idealization Failure in ML

Emily Sullivan

Utrecht University

Idealizations, deliberate distortions introduced into scientific theories and models, are commonplace in science. This has led to a puzzle in epistemology and philosophy of science: How could a deliberately false claim or representation lead to the epistemic successes of science? In answering this question philosophers have been single-focused on explaining how and why idealizations are successful. But surely some idealizations fail. I propose that if we ask a slightly different question, whether a particular idealization is successful, then that not only gives insight into idealization failure, but will make us realize that our theories of idealization need revision. In this talk I consider idealizations in computation and machine learning.

Bridging Scientific Understanding and Creativity with an LLM Benchmark for Narrow-Domain Scientific Fields

Henk W. de Regt and Eugene Shalugin

Radboud University

The rapid advancement of large language models (LLMs) raises questions about their potential to understand complex scientific domains and contribute to scientific discovery. It has been argued that LLMs are 'stochastic parrots' that make predictions on the basis of large training data sets and are therefore incapable of genuine understanding and creativity [2]. However, such claims presuppose certain philosophical conceptions of understanding and creativity, which remain unexamined. We claim, by contrast, that current philosophical insights into the nature of scientific understanding and creativity allow for the possibility of scientific understanding and creativity with LLMs - at least in a restricted sense. Nonetheless, a benchmark for evaluating such understanding and creativity has not yet been developed. The aim of our paper is to fill this gap by developing a semi-automated LLM benchmark creation pipeline for narrow-domain scientific fields for these scientific capabilities, utilizing both human experts and LLMs. With regard to understanding, we adopt Barman et al.'s [1] behavioural conception, which implies that an agent's understanding consists in its ability to process, integrate, and apply knowledge, also in unseen scenarios, beyond mere factual retrieval. In particular, we focus on assessing the type of questions an agent can answer. With regard to creativity, we follow Boden [3] and call a scientific product creative if it is valuable, novel, and surprising.

We propose a pipeline for the semi-automated creation of questions to evaluate LLMs' understanding and creativity in narrow-domain scientific fields. We focus on creating what-, why-, and w-questions (counterfactuals). The questions are generated using human-provided texts with trusted information (e.g. lecture notes, scientific papers, textbooks). Parts of the set of documents are then provided to the LLMs as context with Retrieval Augmented Generation (RAG) and the LLM is tasked with generating questions based on the documents. Domain experts are then invited to review the validity and sensibility of the questions. The resulting question-answer pairs form a high-quality narrow-domain benchmark. We use the field of particle physics as a case study and introduce the platform www.physicsbenchmark.org for generation and curation of high-quality questions. Using this dataset of factual, explanatory, and counterfactual questions, we evaluate how well state-of-the-art LLMs understand particle physics. Contextual grounding is toggled via access to the corpus. We then propose a scientific creativity benchmark (SCB) that challenges LLMs with questions whose answers lie outside their training data. These questions are either unprecedented counterfactual queries or questions about scientific literature published after 2023 (the cutoff date of the tested models). By formulating prompts that fall beyond the statistical patterns the model has learned, we induce a form of novelty: if the model provides a satisfactory answer to, say, a frontier particle physics question, it must have performed non-trivial logical transformations on its outdated information. This deviation serves as an epistemically surprising indicator of creative output, meeting our criteria for responses that are not only novel and unexpected but also valuable. (N.B.: we do not argue that LLMs are creative agents but that they can generate creative products.)

References

[1] K. G. Barman, S. Caron, T. Claassen, and H.W. de Regt, "Towards a Benchmark for Scientific Understanding in Humans and Machines," *Minds and Machines*, 34(1), 2024, doi: 10.1007/s11023-024-09657-1.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.
[3] M. A. Boden, *The Creative Mind: Myths and Mechanisms*. Psychology Press, 2004

Friday, June 27th

What’s Hidden Inside Predictively Successful Deep Learning Models?

Darrell P. Rowbottom

Lingnan University Hong Kong

to be announced

The Underdetermination of Representational Content in DNNs

Sara Pernille Jensen

Oslo University

There is widespread hope of using ML models to make new scientific discoveries. As part of this endeavour, much effort is being put into establishing methods for interpreting the learned basis vectors in the latent spaces of deep neural networks (DNNs) (Räz 2023; Boge 2024), motivated by the belief that the networks implicitly learn scientifically relevant representations or concepts from the data (Buckner 2020; Bau et al. 2020). By studying these learned representations, we may learn about new dependencies and structures in nature. However, there is disagreement regarding how concepts are represented in the hidden layers, specifically whether they are localized or distributed across nodes, and whether they are linear or non-linear.

Here, I argue that for distributed representations, linear or not, the conceptual content of the representations will often be underdetermined. For classical scientific representations, one can unambiguously tell whether a concept has been represented or not, since the concepts represented are, e.g., those explicitly symbolized in an equation. Such representations, with their explicit conceptual content, stand in contrast with mere informational content. For although the information about some derived property is present in the original representation, the derived property is not thereby itself represented.

My worry is whether the different accounts of representations in DNNs lead to any detectable differences between informational and conceptual content in the representations. I assume Harding's operationalized definition of representations in DNNs, requiring a representation to carry informational content about its target, that the later layers of the network use the representation, and that it comes with a possibility of misrepresentation (Harding 2023). Local representations are unproblematic, since each node is dependent on a single variable, so only represents the concept corresponding to that variable. The problem arises for both linear and non-linear distributed representations, where any compound variable derivable from (non-)linear transformations of a set of activations in a given layer may be represented.

Here, the conceptual content will be underdetermined in cases where the concepts include sets of variables which are defined in terms of each other. Examples include the ideal gas law, $PV = nRT$; the total energy, $E_{\text{total}} = E_{\text{kin}} + E_{\text{pot}}$; and the Lagrangian, $L = E_{\text{kin}} - E_{\text{pot}}$. Importantly, these dependencies are often empirical discoveries, not analytic truths, so the concepts themselves are not defined in terms of each other. Yet, when the model represents and conceptualizes two of the three variables, there will be no way for us to tell which of the three that is, due to their interdependence and equivalent model functionality. It will therefore be underdetermined what the conceptual content of the representation is.

I consider two implications of this finding. Firstly, it suggests some caution in our use of such anthropomorphic language of representations and concepts, for if the conceptual content of representations is sometimes underdetermined for DNNs, we might need to reconsider what we really mean by "representations of concepts." Secondly, the underdetermination introduces an additional difficulty in learning new concepts and relations from distributed representations in DNNs, which might have implications for their usefulness in scientific discoveries.

References

[1] Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. "Understanding the Role of Individual Units in a Deep Neural Network." *Proceedings of the National Academy of Sciences* 117 (48): 30071–78. https://doi.org/10.1073/pnas.1907375117.

[2] Boge, Florian J. 2024. “Functional Concept Proxies and the Actually Smart Hans Problem: What’s Special About Deep Neural Networks in Science.” *Synthese* 203 (1): 1–39. https://doi.org/10.1007/s11229-023-04440-8.
[3] Buckner, Cameron. 2020. “Understanding Adversarial Examples Requires a Theory of Artefacts for Deep Learning.” *Nature Machine Intelligence* 2 (12): 731–36. https://doi.org/10.1038/s42256-020-00266-y.
[4] Harding, Jacqueline. 2023. “Operationalising Representation in Natural Language Processing.” *The British Journal for the Philosophy of Science*, November. https://doi.org/10.1086/728685.
[5] Räz, Tim. 2023. “Methods for Identifying Emergent Concepts in Deep Neural Networks.” *Patterns* 4 (6): 100761. https://doi.org/10.1016/j.patter.2023.100761.

The Benchmarking Epistemology – What Inferences Can Scientists Draw from Competitive Comparisons of Prediction Models?

Timo Freiesleben

Tübingen University

Benchmarking—the practice of evaluating machine learning (ML) models based on their predictive performance on test datasets and ranking them against competitors—is a cornerstone of ML research. Often referred to as the "iron rule" of machine learning, benchmarking typically involves four key components: (1) prediction tasks, defining the target variables and predictors; (2) evaluation metrics, determining what constitutes good predictions; (3) datasets, including a training set for model development and a test set for performance assessment; and (4) leaderboards, which rank models based on their test set performance.

As ML becomes a widespread tool in the natural and social sciences, benchmarking is increasingly adopted as a standard method of scientific evaluation [1], [2], [3]. Despite its practical importance, the philosophy of science community has paid surprisingly little attention to benchmarking. Existing discussions have instead focused on inductive inference in ML [4], [5], opacity [6], [7], [8], and explainability [9], [10].

This paper addresses the gap by arguing that benchmarking constitutes a scientific epistemology in its own right, offering a distinct framework for scientific inference. We analyze four types of inferences commonly drawn from benchmarking: (1) identifying the (current) best model for task T; (2) determining the (current) best learning algorithm for tasks similar to T; (3) selecting the (current) most suitable model for deployment in a specific application; and (4) estimating the optimal predictability of a target Y given features X.

A central insight is that none of these inferences can be drawn from benchmark results alone; each requires further assumptions to be valid. Similar to inference from psychological testing, ensuring construct validity is essential [11]. These additional assumptions must be specified and justified by further evidence [12].

To ground our analysis, we examine three case studies from diverse scientific domains: the ImageNet benchmark for image recognition [13], the Fragile Families benchmark for predicting life outcomes [14], and the WeatherBench benchmark for global weather forecasting [15], [16].

Beyond epistemic uses, benchmarks also play crucial social roles—organizing scientific communities around shared goals. However, these social functions can threaten the epistemic validity of benchmarking inferences. For example, iterative use of benchmarks for model tuning undermines the assumption that test data remains unseen, an assumption critical for valid benchmarking [17], [18]. Moreover, benchmarks are often treated as proxies for scientific significance in peer review, incentivizing practices akin to p-hacking [19], [20].

We argue that while benchmarks are powerful scientific tools, they must be used with awareness of their inferential and sociological limitations.

References

[1] Kitchin, R. 2014. *Big Data, New Epistemologies and Paradigm Shifts*. Big Data & Society 1(1): 2053951714528481.

[2] Mussgnug, A. M. 2022. The Predictive Reframing of Machine Learning Applications. *European Journal for Philosophy of Science* 12(3): 55.

[3] Pankowska, P., Mendrik, A., Emery, T., and Garcia-Bernardo, J. 2023. Accelerating Progress in the Social Sciences: The Potential of Benchmarks. `https://doi.org/10.31235/osf.io/ekfxy`.

[4] Sterkenburg, T. F., and Grünwald, P. D. 2021. The No-Free-Lunch Theorems of Supervised Learning. *Synthese* 199(3): 9979–10015.
[5] Karaca, K. 2021. Values and Inductive Risk in Machine Learning Modelling. *European Journal for Philosophy of Science* 11(4): 102.
[6] Creel, K. A. 2020. Transparency in Complex Computational Systems. *Philosophy of Science* 87(4): 568–589.
[7] Boge, F. J. 2022. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines* 32(1): 43–75.
[8] Sullivan, E. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*.
[9] Zednik, C., and Boelsen, H. 2022. Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines* 32(1): 219–239.
[10] Freiesleben, T., Königg, G., Molnar, C., and Tejero-Cantero, Á. 2024. Scientific Inference with Interpretable Machine Learning. *Minds and Machines* 34(3): 32.
[11] AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.
[12] Messick, S. 1995. Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice* 14(4): 5–8.
[13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 248–255.
[14] Salganik, M. J., Lundberg, I., Kindel, A. T., and McLanahan, S. 2019. Introduction to the Fragile Families Challenge. *Socius* 5.
https://doi.org/10.1177/2378023119871580. [15] Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. 2020. WeatherBench: A Benchmark Dataset for Data-Driven Weather Forecasting. *JAMES* 12(11): e2020MS002203.
[16] Rasp, S., Hoyer, S., Merose, A., et al. 2024. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *JAMES* 16(6): e2023MS004019.
[17] Grote, T., Genin, K., and Sullivan, E. 2024. Reliability in Machine Learning. *Philosophy Compass* 19(5). https://doi.org/10.1111/phc3.12974.
[18] Hardt, M., and Recht, B. 2022. *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press.
[19] Gelman, A., and Loken, E. 2014. The Statistical Crisis in Science. *American Scientist* 102(6): 460–465.
[20] Bzdok, D., Altman, N., and Krzywinski, M. 2018. Statistics versus Machine Learning. *Nature Methods* 15: 233–234.

Deep-learning Models and Scientific Understanding through Explanations and Representations

Giovanni Galli

Teramo University

In the rapidly evolving landscape of artificial intelligence, the understandability and explainability of AI systems have become crucial concerns. As AI models grow increasingly complex, they often operate as "black boxes", making decisions without explaining their processes clearly. This opacity can hinder trust, accountability, and ethical compliance, particularly in critical domains such as healthcare, finance, law, and scientific research. Still, deep-learning models (DLMs) are powerful tools in order to understand phenomena, as recognised by Páez (2019), Sullivan (2022), Fleisher (2022), Jumper (2021a) and Abramson et al. (2024). Thus, on the one hand, Explainable Artificial Intelligence (XAI) aims to answer the first issue about the opacity of the DLMs, offering us ways to understand the DLMs; on the other hand, the kind of understanding gained from DLMs leads us to re-define what scientific understanding is.

According to Sullivan (2022), the lack of understanding of DLMs does not limit our scientific understanding of phenomena. She argues that when we fail to achieve understanding with DLMs, it is not due to the lack of understanding of the DLMs in question but to the "link uncertainty", i.e. the lack of evidence, knowledge and understanding of how the model and its target-system are related. On the opposite side, Räz and Beisbart (2022) argue that due to the lack of understanding of DLMs, we may fail to understand a phenomenon scientifically through the use of the models. Along their line, Durán (2021) claims that what we gain from DLMs is not a genuine understanding of the phenomena and that XAI's explanations are better defined as classifications.

In this paper, we first argue that a machine can explain and that some XAI explanations are rule-based. We defend the idea that if specific XAI explanations can capture the rules underlying the scrutinised phenomenon, they are genuinely scientific explanations. Second, we claim that, given understanding as a noetic-mediated state, DLMs play the role of noetic mediators for scientific understanding, even if they present essential differences from other traditionally well-suited mediators, such as explanations, theories, and non-artificial models. Moreover, we highlight a crucial distinction when we speak of scientific understanding with DLMs and with other models and theories. De Regt (2017) and Khalifa (2013, 2017) defend that scientific understanding (SU) is gained via explanatory information about the phenomenon under scrutiny. However, when scientists use DLMs to study and understand a phenomenon they cannot access all the relevant explanatory information. We present the case study of AlphaFold's DLMs (Jumper et al., 2021a, 2021b) to propose another form of SU, complementary to the explanatory one, namely representational understanding (Galli, 2023). We then present the features of representational and explanatory scientific understanding involved in scientific research with DLMs like AlphaFold's models. In conclusion, we outline the differences between representational and explanatory understanding in light of the explanations provided by XAI methods.

Machine Learning, AI, and The Gradability of Explanatory Understanding

Insa Lawler

UNC Greensboro

It is a common place that explanatory understanding comes in degrees. Some people have more understanding of a subject matter than others or a greater degree of understanding. Its gradability is claimed to be one feature that sets apart understanding from knowledge. But what precisely does it mean that understanding comes in degrees or is gradable? In my talk, I explore how the gradability of understanding can be analyzed, drawing on insights from epistemology, formal semantics, metaphysics, and philosophy of science. I will also shed light on what this implies for the understanding we can gain from machine learning models or generative artificial intelligence.

Scientific Progress in the Age of AI

Finnur Dellsén

University of Iceland

What role does artificial intelligence (AI) play – and what role might it play – in scientific progress? Are AI systems best understood as tools for accelerating the scientific progress made by human researchers, or are they capable of making scientific progress in their own right? Could AI systems even become autonomous scientific agents one day, capable of generating, testing, evaluating, and communicating scientific hypotheses in a way that produces scientific progress? And do recent advances in AI research constitute genuine scientific progress, or are these developments better understood as technological advances? This talk explores the relationship between AI and scientific progress through these and related questions, proposing new directions for future research. It explores both how the use of AI in science aligns with or challenges existing accounts of scientific progress, and how the philosophical debate about these accounts sheds light on the value of AI in science.

Predictively-valid “Alien” Features, or Artifacts? Coping with Inscrutable Scientific Progress

Cameron Buckner

University of Florida

Systems like AlphaFold raise the prospect of predictive AI systems that can blow past previous upper bounds on the performance of hand-designed analytical models in many areas of scientific analysis. It is difficult to disagree with the results of these systems, which can achieve predictive accuracy on problems that were thought to be too complex or chaotic for human scientific theory to solve. However, these models may base their predictions on features that are in some sense beyond the cognitive grasp of humans–”alien” properties that may have predictive utility but which are not natural or cognitively accessible to us. In this talk I will analyze these properties by beginning with a discussion of adversarial attacks, and discuss methods for coping with this epistemic situation in a scientific regime which increasingly relies on complex deep learning models for data analysis.

Useful Information

General

Talks will be held at the **Internationales Begegnungszentrum** at TU Dortmund. It is located at Emil-Figge-Straße 59, 44227 Dortmund.

Coffee breaks and **lunch** will be offered at the workshop location. **Wi-Fi** will be available during the conference via eduroam.

Signalchat?

The **conference dinner** will be held at the “Schönes Leben”, at Liebigstraße 23, 44139 Dortmund on Thursday evening at 19:30.

How to get around?

To get to the **conference venue**, you can use the S1 train to the station “Dortmund Universität”. From the station, you’ll need about 5–10 minutes to walk to the ”Internationales Begegenungszentrum”, which is located at Emil-Figge-Straße 59, 44227 Dortmund.

The **NH Hotel** is in the immediate proximity of Dortmund main station. Walk straight to the left when exiting the main station via the front exit; the hotel is within five minutes of walking distance.

The **Schönes Leben** is a 30 minutes walk from the main station or the hotel. You can also reach it quickly from the train station “Möllerbrücke” or the subway station “Saarlandstraße”. Please make sure to keep your tickets when you choose to use public transportation.

Organization and Funding

This workshop is an event organized by the Emmy Noether Group *UDNN: Scientific Understanding and Deep Neural Networks*, generously funded by the German Research Foundation (DFG; grant 508844757). It is also generously supported by the *Department for Humanities and Theology* at TU Dortmund University.