Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 76–87
https://doi.org/10.54941/ahfe1005956

Knowledge Evolution and Scientific Breakthroughs Triggered by AI Hallucinations. A Paradigm Shift?

Anastasia-Maria Leventi-Peetz and Nikolaos Zacharis

Department of Informatics & Computer Engineering, Internet Computing and Cloud Technologies Lab, University of West Attica, Athens, Greece

ABSTRACT

The interdisciplinary impact of artificial intelligence (AI) in science has been especially emphasized by the fact that both, the Nobel Prize in Physics and in Chemistry in 2024 have been awarded for pioneering research with results, decisively based on artificial neural networks. The core of the excelling achievement in chemistry is described as: *capturing of the full computational understanding of living matter at atomic level* (Abriata, 2024). An interesting detail behind this highly acclaimed success, is that one of the laureates had praised AI hallucinations to be the designers of de novo proteins (Anishchenko, 2021). AI hallucinations are defined as incorrect or misleading results, usually produced by models implementing generative AI. Hallucinating AI systems are particularly associated with large language models, chat bots and computer vision tools and their occasionally nonsensical or altogether inaccurate outputs can be welcome in domains such as imaginary and visionary art but they can have significant negative consequences in practical applications. AI systems lack human wisdom. They do not solve problems via understanding context or using ideas of their own. They work with predefined inputs and in the case of generative AI, they generate new patterns some of which may deviate from the knowledge implemented in the algorithm or even defy the wisdom of the algorithm designer. Still, they can prove to be compatible with reality as is the case with de novo proteins. AI hallucinations could then be viewed as glimpses into a future, one yet to be created, for instance when introducing man made proteins and organisms into the existing biosphere. Epistemological questions arising from the perspective that creative mistakes of AI can promote science more effective than human ideas will be discussed. Possible risks in connection to a rapid application of in silico results in structural biology, created mostly with machine learning, will also be considered.

Keywords: Generative artificial intelligence, Computational biology, De novo protein engineering, Biophysics, AI hallucinations, Molecular design, AI reproducibility, Cognitive disruption

INTRODUCTION

Machine learning (ML) and especially *Deep Neural Network* (DNN) techniques are seen as a key instrument in scientific developments, producing results considered as otherwise not possible to achieve (Frueh, 2023). AI is already envisioned as almost qualified to do independent scientific research

as an equal partner at the side of human scientists (Kitano, 2021). According to Kitano (CEO at Sony AI): "AI could grow to implement the science of sciences, in a way that will not resemble the scientific process conducted by human scientists. It may be an alternative form of science that will break the limitation of current scientific practice largely hampered by human cognitive limitation and sociological constraints. It could give rise to a human-AI hybrid form of science that shall bring systems biology and other sciences into the next stage." This statement leaves questions open, especially about the limitations attributed to human intellectual strength. However, some solid worries also arise, associated with the perspective of the realization of a human-machine convergence with all its potential consequences for humanity. In combination with the fact that AI hallucinations, otherwise considered to be AI mistakes or illusions (named after the known for humans condition, due to brain disorders or the influence of drugs) are now considered to be promoting research in biochemistry, increases fears of loosing control. The subject is too big to ever get exhaustively discussed and it grows rapidly bigger, accommodating new issues and results arising on a daily basis. Relevant facts will be outlined and discussed in the next paragraphs.

DATA, INFORMATION, KNOWLEDGE, AI & SCIENCE

Information is the result of processed data that has been organized to get contextual meaning, relevant to some specific purpose, compare Figure 1 (Cotton, 2023). Though information and knowledge are related, they still are different conceptions. It is more than information needed to acquire knowledge, because knowledge is also awareness, insight and understanding, gained through experience, education and analysis. Knowledge encompasses the use of information and can involve personal and collective expertise. There exist various kinds of knowledge, including explicit, implicit, tacit, declarative, procedural etc. Implicit and tacit knowledge are associated with personal experience, practice and skills of individuals, they involve intuitive understanding and insights developed over time and are difficult to articulate or document. Not all kinds of knowledge can get implemented in data. There is no consensus on whether AI produces information or knowledge. The prevailing opinion is that AI produces information rather than knowledge, but there are also opinions that both information and knowledge can be produced. Knowledge involves understanding the meaning and implications of information and entails the capability to apply this understanding to new situations. With their predictive capacities, AI models can mimic aspects of knowledge but the process lacks deeper understanding and contextual awareness that characterize human knowledge. If AI models could acquire or produce knowledge, they would be able to generalize from training data to make predictions on new, unseen data. Unfortunately, this generalization capability is missing in the majority of real-world cases. Can AI produce *scientific* results? Scientific results must adhere to principles and criteria that ensure rigor, objectivity and reliability. The results must be based on empirical evidence, involving careful observation or experimentation, they

must be reproducible, evaluable through peer review, generated by a sound and comprehensible method which allows their replication under the same conditions, thereby confirming their validity and credibility. This means, the results have to be verifiable and falsifiable, it should be possible for them to be refuted by evidence, observation or experiment. The results must also establish causal relationships which is a key aspect of scientific knowledge. Finally, they must be logically consistent with the established body of scientific knowledge. Any anomalies or inconsistencies must be thoroughly investigated and explained. In this sense, is the expectation that AI models can be a source of leading-edge scientific ideas and results well-founded?

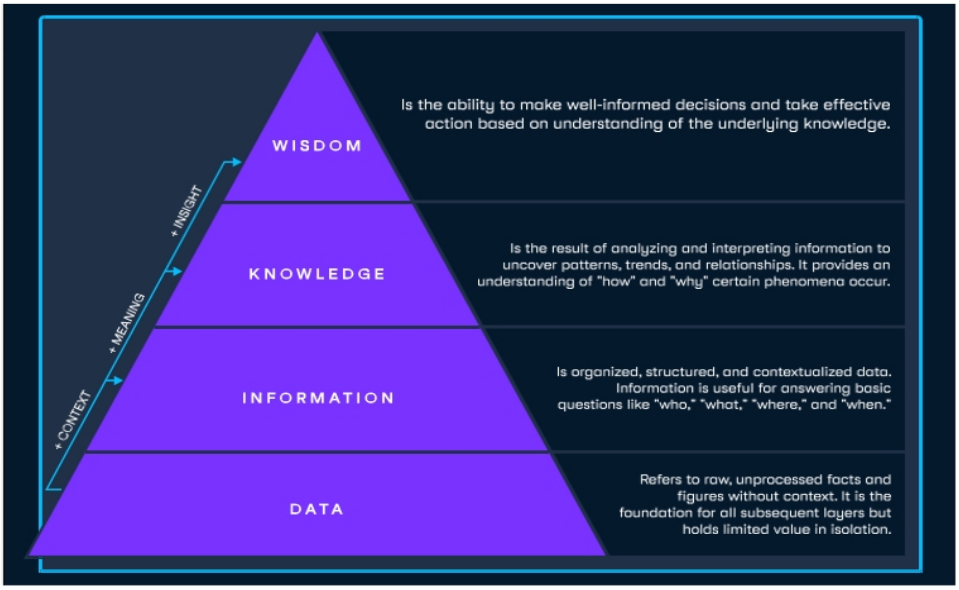


Figure 1: The *Data-Information-Knowledge-Wisdom* (DIKW) pyramid illustrates the progression of raw data to valuable insights. It gives a framework to discuss the level of meaning and utility within data. Each level of the pyramid builds on lower levels, and to effectively make data-driven decisions, it takes all four levels (Cotton, 2023).

REPRODUCIBILITY OF SCIENTIFIC ML-RESULTS

A recent study, conducted at the Department of Computer Science at Princeton University, focuses especially on reproducibility issues in science based on ML, which involves making scientific claims using results of ML models as evidence. In May 2024, an update of a running list of papers with reproducibility failures or pitfalls in ML based science, has been published (Kapoor, 2024). This list contains 41 papers from 30 scientific fields where errors have been found, collectively affecting 648 papers and in some cases leading to wildly overoptimistic conclusions. Medicine, molecular biology, clinical epidemiology, neuroimaging, genomics and pharmaceutical sciences are some of the listed fields. The so-called *data leakage* in data analysis has been identified to be a leading cause of errors. This leakage occurs when the

training dataset overlaps with the test data, or when a model is trained on data that would not be available in a real deployment scenario and can lead to inflated model performance metrics. The ML community has investigated the impact of leakage in several engineering applications and mitigation strategies have been suggested, however, data leakage occurring in ML based science has not been comprehensively investigated. Specific challenges, causes, and effects within scientific applications of ML have not been thoroughly studied. In many scientific fields where ML is applied (such as biology, physics, or social sciences), data may have unique characteristics or structures that can introduce new or subtle forms of leakage. Specialists call it a crisis for two reasons: First, reproducibility failures in ML based science are systemic. In nearly every scientific field that has carried out a systematic study of reproducibility issues, the majority of the reviewed papers suffered from these pitfalls. Second, despite the urgency of addressing reproducibility failures, there are not yet any systemic solutions.

HOW SCIENTIFIC IS DEEP LEARNING?

Although there exist theories about how NN architectures and training methods will perform, and mathematical frameworks provide the according foundations, there is no universal recipe for the adjustment of parameters of DNN models. The optimal configuration depends on the dataset, the model architecture and the task at hand. One has to rely on experimentation to find what works best, which is mostly not straightforward, comprehensible or transparent. Despite advances in automated hyperparameter optimization, a lot of hyperparameter tuning still relies on trial and error. The process is time-consuming and resembles rather *dark art* than scientific routine. The landscape of the loss function of a DNN is usually highly non-convex, containing numerous local minima and saddle points. A DNN model can converge to different local minima, which may not correspond to the globally optimal solution. AI models, especially those with billions of parameters, can be considered underdetermined or overparametrized in many contexts because they have typically more parameters than constraints, which allows that multiple configurations can yield similar performance on the training data. Underdetermined systems have many potential solutions which entails many challenges. For instance, ambiguity in picking the most appropriate for a given situation solution. Or lack of stability, meaning that the model might find a solution that fits the training data well but doesn't generalize to other datasets. Also, a minor, imperceptible to humans perturbation in the input data, can cause incorrect predictions. The lack of stability of such models makes them vulnerable also to adversarial attacks. Moreover, with many possible solutions, it can be challenging to interpret and understand how the model reached a particular solution, which complicates debugging and validation processes. ML models, big or small, are also susceptible to the Rashomon Effect which refers to the phenomenon that multiple models or algorithms, even if they are applied to the same data and task, may produce different outcomes or predictions, each seemingly valid in its own right (Marx et al., 2020). The collection of these different models or

solutions is called *Rashomon set*. The cornerstone of any effective AI model is the ability to generalize well which ensures that a model performs well on new, unseen data. Unseen data, in the context of AI models, typically refers to data that was not part of the training set. It is normally assumed that the data in both the training and the test set are drawn from the same probability distribution. In this case the data is said to be IID, or *Independent and Identically Distributed*. If the test data does not share the same statistical properties with the training data, then the data is *Out of Distribution* or OOD. In such cases, the model usually does not perform as expected. In many realworld applications, a model encounters OOD data. Different environment conditions, the adding of new types of data features, the involvement of human behavior in domains like social media, marketing etc. can lead to unpredictable factors influencing the stability of the data distribution. In healthcare applications, new diseases, new treatment protocols, or patient demographics can often result in OOD data. In the financial sector, market conditions, economic events, and regulatory changes can introduce OOD data. Of course, other model properties like accuracy, scalability, interpretability, fairness etc. are also very important, however, the property of generalization is foundational for achieving the other properties, as there are dependencies. Unfortunately, not only generative AI models produce hallucinations, compare Figure 2 (Bhadra et al., 2021; Jabbour et al., 2023).

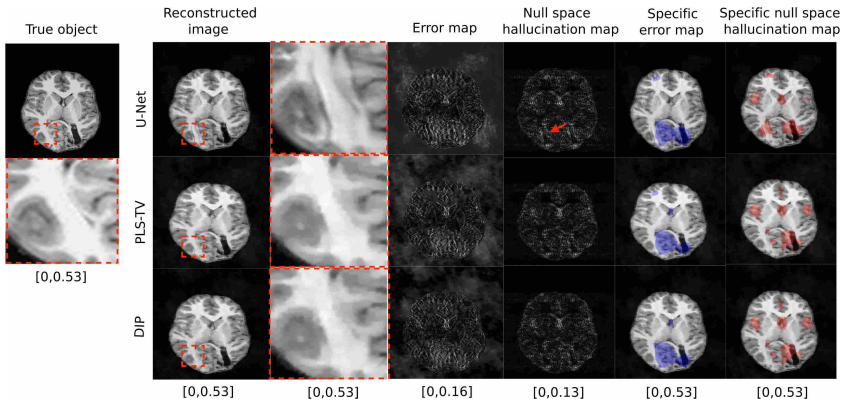


Figure 2: Hallucinations in tomographic image reconstruction. True object and reconstructed images with error map and hallucination maps for OOD data with different reconstruction methods: U-Net (top), PLS-TV (middle) and DIP (bottom). The image estimated by the U-Net method has some distinct false structures (region within red bounding box) that do not exist in the reconstructed images obtained by using PLS-TV and DIP. This region is also highlighted in the specific null space hallucination map for the U-Net, a convolutional NN model, which indicates that the false structure is a NN hallucination (Bhadra et al., 2021).

DEEP LEARNING & CROSS DOMAIN SCIENTIFIC RESULTS

Cross-domain combination of scientific results using ML, is already widely employed and looks quite promising. An example is the combination of material science with computational molecular biology to develop new drugs

(Jeewandara, 2023; Wang et al., 2025). A major challenge in drug discovery is protein structure prediction and ML models significantly advance the analysis of target proteins, promoting the design of drugs that interact more effectively with them. AlphaFold2, for example, uses ML to predict protein 3D structures and has revolutionized drug design. However, there are serious challenges that have to be considered: Many complexities are inherent in both biological systems and materials. In biology, data can be noisy, incomplete, or inconsistent. For example, genomic, proteomic, and metabolic data may have missing values or experimental errors, making it difficult for ML models to draw accurate conclusions. Data used in material science, such as properties of materials, synthesis conditions or experimental results, may not always be comprehensive or standardized. This creates some serious issues when trying to predict or simulate properties of new materials accurately. Biological systems are multifaceted as at least five different layers of biological information flow and processes that contribute to their functioning have been found (Hasin et al., 2017). Together, the layers capture both the static blueprint (genetics) and the dynamic, context-dependent processes that influence an organism's behavior, function, and health. The information layers are not yet completely understood. Each layer operates at different spatial, organizational and multilevel temporal scales, and layers influence each other in a highly dynamic manner. The data of the layers have different types, they may not be compatible or require sophisticated preprocessing and modeling techniques to get integrated into cohesive ML models that can predict outcomes such as drug efficacy or side effects. Models to predict drug efficacy, combining molecular properties with biological activity data, have demonstrated exceptionally good performance on training datasets but failed to generalize to new molecules due to overfitting. They "memorized" specific patterns in the training data rather than learning generalizable relationships. Without clear insight into how a model arrives at its predictions, it can be challenging to explain its reasoning or ensure that the drug is genuinely targeting the right biological pathways or disease mechanisms. The problem of models that produce in silico promising results but fail to anticipate safety concerns, leading to toxicity, organ damage, or severe side effects in clinical trials is well known. Many discrepancies arise mainly due to the complexity of biological systems which are difficult to fully capture in computational models (Simon, 2013; Aliferis, 2024; Wang, 2022; Mann, 2021; Dara, 2022).

AI HALLUCINATIONS & PROTEIN DESIGN

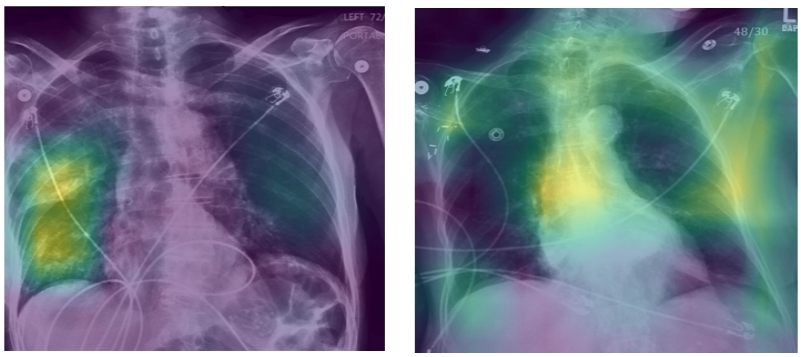
Generative AI models are designed to create new content that resemble patterns from the data they were trained on. A famous example of generative AI is the GPT (*Generative Pre-trained Transformer*) series of *natural language processing* (NLP) models, which can generate human like text, based on inputs received. It is used for various applications, including chatbots, content creation, translation etc. It is almost impossible to create reproducible and consistent results using generative models. AI hallucinations refer to cases of AI models generating outputs or predictions, that are either incorrect,

unrealistic, or entirely fabricated. In the context of NLP, hallucinations often refer to text which sounds plausible but is factually incorrect or made up. Models operate on the basis of the most statistically likely sequence of words according to their training data and may combine text pieces in creative ways leading to unexpected or false connections. Hallucinations, typically considered a negative aspect in AI development, can be misleading with harmful consequences, especially in sensitive applications like healthcare, legal analysis etc. but they are seen as having a positive or inspirating potential in abstract art, entertainment, imaginative writing etc. David Baker, along with other researchers, was part of a groundbreaking achievement in protein design that ultimately contributed to their winning the Nobel Prize in Chemistry in 2024. While the full details of the Nobel-winning research are still unfolding, the key aspects of how the deep learning software tools RoseTTAFold and trRosetta played a role in this achievement are connected to advancements in de novo protein design and structural biology. The RoseTTA family of software tools has been developed over years by researchers at the University of Washington's Institute for Protein Design, led by David Baker. Baker's latest innovation consists of a "three-track" neural network architecture that simultaneously considers patterns in protein sequences, amino acid interactions, and possible 3D structures. "AI hallucinations were central to making proteins from scratch", said Baker, adding that they helped his lab to design around 10 million "all brand-new proteins that don't occur in nature". And he concluded: "Things are moving fast. Even scientists who do proteins for a living don't know how far things have come" (Smith, 2024). It seems that the inventors of these new ingenious methods feel themselves somehow overwhelmed by their own achievements. According to the related publications, the trRosetta DNN is advanced to the point that it can consistently predict protein structure quite accurately for de novo proteins design, from just a single sequence, without using co-evolution information. The information stored in the many parameters of the network make it capable to generate physically plausible backbones and amino acid sequences which encode them. With this statement, the developers attribute generative AI properties to trRosetta. Baker's team took 100 randomly created amino acid sequences and fed them into the trRosetta network. The total number of the resulting *hallucinated* proteins is not explicitly given in the description of the experiment. However, 129 of them were checked with independent folding simulations and it could be confirmed that the generative network produced sequences that indeed encode the corresponding structures. 27 of the 129 proteins could get also experimentally validated in various ways. It was so demonstrated, that a DNN, trained exclusively on native sequences and structures, generalized to create new proteins whose sequences are completely unrelated to those of native proteins and which fold into stable structures. The authors underline "the power of generative deep learning approaches for molecular design which will undoubtedly continue to grow over the coming years."

UNIQUENESS AND STABILITY OF PROTEIN PREDICTIONS?

One of the biggest challenges in accurately predicting proteins and their interactions, is the flexibility of proteins. Proteins, mostly visualized as fixed 3D formations, are in reality very flexible and highly dynamic molecules that undergo significant conformational changes, adopting different structural states, depending on their environment, interactions, and functional needs. While tools like AlphaFold2 and Rosetta are extremely powerful for predicting static structures, they have limitations when it comes to predicting the dynamic behavior of proteins. Protein models are trained on datasets from the *Protein Data Bank* (PDB), the primary repository of already experimentally analyzed protein structures. Small, stable and well-ordered protein conformations are mainly represented in the PDB, whereas important protein classes such as, large, complex, flexible and membrane proteins, are significantly underrepresented. This is also due to challenges associated with their experimental analysis using X-ray crystallography or *nuclear magnetic resonance* (NMR) spectroscopy. There exist conformational states that are transient, they are rare, or occur under specific conditions, making them difficult to observe experimentally or predict computationally. Proteins adopt also conformations that are energetically less favorable and therefore exist in very low populations. These states have enormous functional importance due to their specialized roles in many essential biological processes and are considered to have critical and sometimes life-saving roles in cellular and organismal health. Low-population conformations are particularly important in the context of drug discovery, especially because some drugs are designed to stabilize or promote specific conformations, including those that are less stable. Proteins can adopt unique conformations in specific environments, such as within a crowded cellular milieu or under stress conditions, that are not replicated in experimental setups. Models oversimplify the environment which can result in predictions that are far from reality. Many important proteins are *intrinsically disordered* (IDP) or contain disordered regions which are crucial for a variety of essential biological functions, like cell signaling, gene regulation, protein-protein interactions and molecular recognition. Misfolding or aggregation of IDPs is often linked to diseases like Alzheimer's, Parkinson's, and Huntington's disease, therefore, understanding them is critical in molecular biology. AlphaFold2 cannot predict the multiple conformations or dynamic transitions that IDPs undergo. RosettaFlex, developed to predict disordered proteins, has also limitations and it cannot simulate the full dynamic behavior or time evolution of IDPs. Proteins can undergo spontaneous conformational fluctuations that happen on multiple timescales, ranging from picoseconds to seconds or longer. If they are not fully accounted for in models, inaccurate predictions can be produced. It is computationally expensive to simulate adequately long timescales to fully capture interaction-relevant conformational changes. Simulations typically sample only a limited portion of the conformational space, which can lead to incomplete or biased predictions. Failures to predict the efficacy of molecules in inhibiting proteins in real-world experiments have been attributed to this issue. Models for drug discovery often focus on

interactions between a drug and its intended target protein, but they may fail to predict off-target interactions. It is a problem when a drug binds to unintended proteins, which can lead to adverse effects that were not foreseen by the model. Also, the relationship between amino acid sequence and protein structure is not always one-to-one. This is known as *sequence-structure degeneration* and refers to the fact that a protein's final shape is not entirely deterministic. For some proteins, there may be more than one *correct* and stable conformations, which can influence how they interact with other molecules or perform their biological functions. The amino acid sequence also does not encode *post-translational modifications* (PTM) of proteins which can influence their structure and functionality. In biological environment, proteins may undergo modifications, which can stabilize or destabilize them, depending on environment context. Experimental stability measurements are often conducted on the unmodified protein (or with specific PTMs added in an engineered way), which may not fully capture how the protein behaves in vivo. In living organisms, proteins are also subject to degradation processes. Even if a protein is stable in vitro, it might be degraded rapidly in a biological system if recognized as misfolded, damaged, or lacks necessary signals for stabilization (such as chaperone interactions). Oxidation can also lead to a loss of structural integrity or functional activity, affecting the protein's stability. A biological environment is often more oxidizing or reducing than the conditions used in many in vitro stability experiments. Neither AlphaFold2 nor trRosetta are designed to calculate or predict protein degradation. In vivo stability, functionality, and behavior often need to be tested separately, even after in vitro stability calculations and experimental validation (Yang et al., 2020).



(a) (b)

Figure 3: AI assisted diagnosis of respiratory failure with explanation (heatmap), based on real clinical X-ray vignettes of patients. Heatmaps show which part of the vignette the AI model was attending, when making its diagnosis. When the AI diagnosis is correct (a), adding the explanation increases the clinicians' accuracy by 4.4% as compared to 2.9% without heatmap. An intentionally biased AI model in (b) highlighted features of age in a patient's chest, such as bone density, and wrongly diagnosed pneumonia. Though obviously irrelevant features and partly also outside of the lungs were highlighted, some clinicians still diagnosed the patient with pneumonia (Smith, 2023; Jabbour et al., 2023).

CONCLUSION

Artificial intelligence has undeniably undergone spectacular development in recent years, with groundbreaking advancements that have transformed numerous industries and everyday life. The issues discussed above may perhaps temper uninhibited willingness to accept all AI results prematurely and uncritically. There is a tendency for algorithmic monocultures to prevail in ML, with an over-reliance on large models. As already discussed, large models have a greater capacity to overfit data. This makes them less generalizable and prone to poor performance on unseen data, which is particularly problematic in fields such as healthcare, bioinformatics, finance etc., where errors can have significant consequences. Large models are particularly known for producing *spurious correlations* and *shortcuts* (Geirhos et al., 2020). Their training is prohibitively expensive for many individual researchers, especially those in smaller institutions or developing countries. This reduces the diversity of perspectives in the development of AI. It could end up in a situation where only certain groups, usually those with the financial means or commercial interests, drive the research agenda and valuable insights from under-represented communities may be missed. The uncritical over-reliance on large, opaque models, that are considered *state-of-the-art* in many application areas, could have unforeseen consequences. For instance, biased predictions from dominant models may be taken as unquestionable truth because the models achieve high accuracy and alternatives don't exist. Many examples show that the accuracy of a prediction does not guarantee its correctness. The vast computational resources, needed to train large models also contribute to significant environmental costs, both in terms of energy consumption and carbon emissions, which is a critical concern given the global need for sustainability. Efforts to develop smaller, more transparent models, that can perform as well as larger ones, have proven fruitful and continue to be useful, as Cynthia Rudin, the leading advocate of interpretable machine learning, has shown (Rudin et al., 2022). In her research, Rudin has demonstrated that smaller and simpler models can achieve comparable performance to complex models in many domains. This means, interpretability does not necessarily come at the expense of performance. Interpretable models are easier to audit for bias and fairness, two concerns that have become increasingly important in AI development. When a model's decision-making process is transparent, it is easier for users to understand and trust its recommendations to make informed decisions. Transparent models represent human-centered design. One more advantage of interpretable models is that they are often more amenable to open-source development. Models are easier to implement, explain, and share. This fosters greater collaboration and access to cutting-edge research, enabling a wider community of researchers to contribute to AI advancements. Rudin's work calls for a balanced approach, one where interpretability is prioritized alongside performance. This shift could have significant positive impacts on the accessibility, fairness, and ethical deployment of AI across a wide range of domains.

Data, when presented in large quantities and processed through AI, often carries an air of objectivity and infallibility. Conclusions drawn by AI systems can be thought of as ground truth, leading to a devaluation of subjective human experience and reasoning. Becoming passive recipients of decisions and actions reduces the engagement with one's own critical thinking, it can lead to mistakes (Smith, 2023) and to a sense of losing control over one's life. The more people rely on data and *black box* models to make decisions for them, the more human agency and a sense of personal responsibility could be eroded. The belief that AI will always provide the most relevant and accurate information could limit opportunities for critical thinking and expanding knowledge, as people may stop seeking out different points of view. Technological disruptions caused by spectacular developments in AI, such as protein design by AI hallucinations, could also lead to cognitive disruptions. This can happen if technologies are created that scientists themselves don't fully understand or control. While synthetic biology holds incredible promise for addressing global challenges in healthcare, environmental sustainability, and food production, it also poses significant risks that could have negative and irreversible consequences for humans and the environment. These risks include unintended health effects, the creation of harmful pathogens, ecological disruption, and ethical dilemmas related to human enhancement and reproduction. Careful regulation, robust ethical frameworks, transparent oversight, and international cooperation are essential to ensure that technologies are developed and applied in ways that prioritize human safety, social equity, and environmental sustainability. Maintaining human-centered thinking is critical, as is cultivating an awareness of AI's limitations and ensuring that AI augments rather than replaces, human cognition, compare Figure 3. Fostering a healthy relationship with technology, can help to avoid getting overwhelmed by it and preserve the ability to think critically and meaningfully.

REFERENCES

Abriata, L. A. (2024) *The Nobel Prize in Chemistry: past, present, and future of AI in biology*, Commun Biol, Volume 7, doi: 10.1038/s42003-024-07113-5.

Aliferis, C. et al. (2024) *Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI*. In: Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Springer, doi: 10.1007/978-3-031-39355-6 10.

Anishchenko, I. et al. (2021). *De novo protein design by deep network hallucination*. Nature 600, doi: 10.1038/s41586-021-04184-w.

Bhadra, S. et al. (2021) *On Hallucinations in Tomographic Image Reconstruction*, IEEE Transactions On Medical Imaging, Vol. 40(11), doi: 10.1109/TMI.2021.3077857.

Cotton, R. (2023) *The Data-Information-Knowledge-Wisdom Pyramid*, https://www.datacamp.com/cheat-sheet/the-data-information-knowledge-wisdom-pyramid.

Dara, S. et al. (2022) *Machine Learning in Drug Discovery: A Review*, Artif Intell Rev, Vol. 55, doi: 10.1007/s10462-021-10058-4.

Frueh, S. (2023) *How AI Is Shaping Scientific Discovery*, https://www.nationalacademies.org/news/2023/11/how-ai-is-shaping-scientific-discovery.

Geirhos, R. et al. (2020) *Shortcut learning in deep neural networks*. Nat Mach Intell, Vol. 2, doi: 10.1038/s42256-020-00257-z.

Hasin, Y. et al. (2017) *Multi-omics approaches to disease*, Genome Biology Vol. 18(83), doi: 10.1186/s13059-017-1215-1.

Jabbour, S. et al. (2023) *Measuring the Impact of AI in the Diagnosis of Hospitalized Patients. A Randomized Clinical Vignette Survey Study*, JAMA, Vol. 330(23), doi: 10.1001/jama.2023.22295.

Jeewandara, T. (2023) *Regenerative medicine at the intersection of materials science and biology*, https://communities.springernature.com/posts/regenerative-medicine-at-the-intersection-of-materials-science-and-biology.

Kapoor, S. and Narayanan, A. (2024) *Leakage and the reproducibility crisis in machine-learning-based science*, Princeton University, USA, https://reproducible.cs.princeton.edu.

Kitano, H. (2021) *Nobel Turing Challenge: creating the engine for scientific discovery*, NPJ Syst Biol Appl, Vol. 7(29), doi: 10.1038/s41540-021-00189-3.

Mann, M. et al. (2021) *Artificial intelligence for proteomics and biomarker discovery*, Cell Systems, Vol. 12(8), doi: 10.1016/j.cels.2021.06.006.

Marx, C. T. et al., (2020) *Predictive Multiplicity in Classification*, preprint doi: 10.48550/arXiv.1909.06677.

Rudin, S. et al. (2022) *On the Existence of Simpler Machine Learning Models*, ACM Conference on Fairness, Accountability, and Transparency, doi: 10.1145/3531146.3533232.

Simon, R. et al. (2013) *Overfitting in prediction models – Is it a problem only in high dimensions?*, Contemporary Clinical Trials, Vol. 36(2), doi: 10.1016/j.cct.2013.06.011.

Smith, D. (2023) *Clinicians could be fooled by biased AI, despite explanations*, https://news.engin.umich.edu/2023/12/clinicians-could-be-fooled-by-biased-ai-despite-explanations/.

Smith, D. (2024) *Scientist says the one thing everyone hates about AI is ultimately what helped him win a Nobel Prize*, https://fortune.com/2024/12/24/ai-hallucinations-good-for-research-science-inventions-discoveries.

Wang T. et al. (2025) *Advanced Deep Learning Methods for Protein Structure Prediction and Design*, preprint: doi: 10.48550/arXiv.2503.13522.

Wang Z. et al. (2022) *Artificial Intelligence for In Silico Clinical Trials: A Review*, preprint: doi: 10.48550/arXiv.2209.09023.

Yang, J. et al., (2020) *Improved protein structure prediction using predicted interresidue orientations*, Biophysics and Computational Biology, Vol. 117(3), doi: 10.1073/pnas.1914677117.