CS 562 Assignment 1

Reflection on Mariette Awad's paper: 'Types of AI and their use in Science'

Question 1:

Awad distinguishes between different "types" of AI. What classification scheme does the paper use and why do these types matter for scientific research?

Answer:

Classification scheme:

In my opinion, Awad took an **application-driven scheme** when defining the 'types' of AI, rather than focusing on the underlying technical architectures (e.g. supervised vs unsupervised) commonly used in the realm of computer science.

Awad grouped AI into 7 types according to how each type is **applied** in a scientific context:

- Predictive AI anticipates outcomes and models complex systems
- Descriptive AI uncovers patterns in large datasets
- Generative AI produces new content and hypotheses
- Optimization AI automates and refines experimental processes
- Causal and interpretable AI explains how and why results are reached
- Privacy-aware AI enables collaboration on sensitive data
- Meta-scientific AI supports the scientific process itself by generating hypotheses and connecting insights across fields

It is worth noting that this application-driven approach tells us more about what purposes AI serves than about AI itself. A computer scientist might draw very different boundaries, grouping these same tools by architecture or learning paradigm, however, given the paper's explicit goal of informing the broader scientific community, such grouping by application area is a more appropriate choice in my opinion.

Why do the types matter for science:

This classification matters for scientific research because it makes clear how to pick your AI tool as a scientist when it comes to different use cases of the scientific research. For example, a researcher designing an experiment may need meta-scientific AI; one trying to make sense of a model's output needs causal or interpretable AI; another one working with sensitive medical data needs privacy-aware AI. By framing AI in the context of scientific applications, Awad makes the potential of AI clearer to her intended audience of scientists

and policymakers, who need to understand what AI can **do** for their work, rather than how it works internally.

Question 2:

Does Awad make a clear distinction between AI as a tool and AI as a scientific collaborator? If so, what are the differences and what are some examples given to support the differences? Do these examples suggest a real shift in how science is conducted, or mostly an extension of existing methods?

Answer:

Does Awad make a clear distinction:

In my opinion Awad made a meaningful and clear distinction between AI as a tool and AI as a scientific collaborator. The difference was mainly defined by the degree of autonomy and 'epistemic' contribution the AI system brings to the research process.

What are the differences and supporting examples:

When Awad described AI as a **tool**, it operates as a sophisticated instrument directed mainly by human researchers. A clear example would be AlphaFold, which predicted over 200 million protein structures from amino acid sequences, dramatically accelerating a task that traditionally required resource-intensive techniques like X-ray crystallography. Another example is DeepSEA, which uses CNNs to predict the regulatory effects of noncoding genomic variants. Both systems perform a specific, human-defined task at a much bigger scale and higher speed than humans, while the scientific questioning, the design of hypotheses, and the interpretations all remain with the human researchers.

On the other hand, when Awad described AI as a **collaborator**, she pointed to systems that can generate hypotheses, design experiments, and connect insights across disciplines with some degree of autonomy. **Section 1.7** on meta-scientific AI describes systems like Google Co-Scientist and SciAgents as capable of "identifying unexpected relationships, refining theoretical models and accelerating materials discovery". These functions are beginning to resemble the reasoning of a junior researcher rather than a complex calculator. The paper's use of the term "epistemic agent" essentially summarizes this distinction: an epistemic agent doesn't just compute answers, it potentially influences what questions get asked and what counts as evidence.

Do the examples suggest real shift or just an extension:

In my opinion, the examples suggest a **real but rather early-stage shift**. For well-defined tasks like protein folding or genomic analysis, AI is extending existing methods with great efficiency. For open-ended discovery, the collaborator mode is compelling but still looks to be in its early stage and largely aspirational. The shift in how science is conducted is beginning but has not fully arrived yet.

In the paper, many of the meta-scientific systems Awad mentioned still depend heavily on human-defined goals. As long as humans are directing the AI's inquiry and validating its outputs, the boundary between a very powerful tool and a true collaborator remains blurry. Like the Galactica case where an LLM trained on millions of scientific papers was withdrawn after producing fabricated citations, it illustrates precisely the gap between seemingly autonomous output vs reliable epistemic evidence.

Question 3:

What are some limitations or risks of using AI in science? How do these relate to issues such as interpretability, bias, reproducibility, or theory formation?

Answer:

Awad acknowledges several significant limitations and risks of using AI in science.

Firstly, the most structurally threatening to science in my opinion is the '**black box' opacity** problem, inherent in deep learning models. To me, science as a discipline is fundamentally built on transparency and replicability, for which researchers must be able to understand, communicate, and reproduce the reasoning behind a finding. Deep learning models could violate this principle by producing outputs through processes that resist human interpretation. This is not merely inconvenient; it undermines the peer review process and makes it difficult to distinguish genuine scientific insight from sophisticated pattern-matching.

Awad addresses this through causal and interpretable AI in Section 1.5, pointing to techniques like SHAP and LIME as tools for making black-box models more transparent. However, these solutions are limited. SHAP and LIME are essentially wrappers over the underlying opacity — they approximate explanations rather than reveal true reasoning, and Awad's paper acknowledges that their outputs "can vary depending on the model and data used." An explanation that is inconsistent across runs or sensitive to data inputs should not be considered as a reliable scientific account. The black-box problem is therefore only mitigated but far from resolved.

A second significant risk mentioned by Awad is **bias**, specifically with NLP tools. The paper cites Caliskan et al. (2017) to show that NLP models can perpetuate racial and gender stereotypes derived from training data. In scientific contexts where NLP is used for tasks such as assessing public opinion, or mining literature, embedded bias can systematically skew what questions get foregrounded and whose knowledge gets recognized, threatening the supposed objectivity of science.

Awad also mentioned about **hallucination in LLMs**, which in my opinion poses a direct threat to theory formation and reproducibility. The Galactica case is a clear example: a model trained on scientific papers itself produced fabricated citations and misleading claims, demonstrating that fluency and apparent authority are no guarantee of accuracy. If AI-generated hypotheses or literature summaries enter the scientific record without rigorous validation, the cumulative knowledge base is at risk of corruption, and this could significantly hurt theory formation.

Question 4:

According to Awad's arguments, is AI more likely to accelerate scientific discovery or to reshape the scientific method itself? Do you agree or disagree?

Answer:

Awad's central argument leans toward AI **reshaping the scientific method itself** rather than merely accelerating it. In the closing thoughts section, Awad cited Thomas Kuhn's concept of paradigm shifts, where she frames AI's integration into science as a fundamental reshaping of how we should carry out science in the future. However, I do not personally agree with her standpoint, here is why:

When examined closely, the paper's own evidence tells a more modest story. The most concrete examples such as AlphaFold predicting protein structures, DeepSEA analyzing genomic sequences, and NeuralGCM improving climate forecasting, are all cases of AI performing human-defined tasks with greater speed and scale. In each case, humans set the scientific question, curated the training data, and interpreted the results. These are powerful extensions of existing methods, not replacements of the scientific method itself. The meta-scientific AI systems in Section 1.7, which came close to the 'AI-as-a-collaborator' framing, sounded more aspirational than evidence-based in my opinion, and it was definitely not at the same level of substantiation compared with the standpoint of 'AI-as-a-tool'.

There is therefore a **meaningful gap between Awad's argument and her evidence**. Furthermore, the limitations Awad herself identifies make it difficult to accept the reshaping

claim in good faith. The black-box opacity problem, which she acknowledges is only partially addressed by techniques like SHAP and LIME, means that AI-generated knowledge currently lacks the transparency that scientific reasoning requires. If we cannot reliably explain how an AI reached a conclusion, it cannot yet serve as a genuine epistemic partner in reshaping how science is done.

On a balanced note, to me AI is most accurately described at this stage as **accelerating** scientific discovery **within** the existing methodological frameworks. The potential to reshape the scientific method still largely unrealized.