Research Statement ~ *Austin Clyde*

Public trust in science is declining, and new technical systems outpace regulatory environments and legal sense-making. Artificial intelligence (AI) systems produce uncertainty and opacity in basic scientific research, medical practice, and criminal sentencing. How do we, as scientists and citizens, understand the results of AI systems, the reasoning behind AI systems, and their role in decision-making? Understanding, scrutiny, public trust, regulation, and the rule of law are all questions of *interpretation*. Interpretation is the study of meaning and sense-making—how is it we go from some "full" understanding to raising a question about some part of it, acting on the question, and reflecting on and sharing a reinterpretation of our original understanding? Questions of interpretation are often excluded from computer science domain research since the scientific method is assumed to be free from social interpretation. But as AI, big data, and high-performance computing (HPC) reshape the scientific method and become entangled in exercising power throughout the daily lives of global citizens, we must understand the landscape of possibilities for interpreting our usage of AI systems, interpreting the results of these systems authentically and faithfully, and designing scientific and practical systems which are open to plural, flexible, and democratic means of interpretation.

My research interest is developing interpretable AI for science systems and analyzing the role of interpretation and power in algorithmic decision systems. My research is both quantitative and qualitative. Motivated by my research in computational drug design, I aim to develop computational methods for increasing **scrutiny** and **calibration** with large language models (LLMs) in scientific applications. Second, motivated by current challenges in explainable-AI (X-AI) research, I aim to build data-based **explainability** techniques that incorporate plural understanding by focusing on the causal relationship between data included in the training data and the inferences a model makes. Third, my work is analytical to understand better how different groups interpret models in practice in a **democratic** context. I use methods from science and technology studies (STS) to analyze the institutional and social structures that perpetuate kinds of interpretation, such as how these methods stabilized as opposed to alternatives. This work profoundly relates to deliberative democratic theory, how pluralistic understanding of a system can be fostered, and to legal studies, where the interpretations of constitutional or human rights are being reapplied and reconsidered considering new technical developments. The impact of my work can be viewed broadly through the lens of opening the black box of accelerating technology development to scientists, policymakers, and citizens alike. This research can also be regarded as an alternative approach to "AI ethics" by understanding the social impact and "ethical" impacts of AI through empowering reflection rather than following a particular narrative of AI in light of what one decides is "virtuous" or "ethical."[1] By focusing on studying our interpretive relationship with scientific tools, algorithms used in decision-making, and law, my work is central to increasing civic engagement with technology and channeling technology through democratic values.

Area 1: Explainability. Explainability is essential to the current regulatory landscape of AI, human-computer interaction, and accelerating scientific discovery with AI and HPC; however, recent explainability research focuses on models *as decision makers*—i.e., explainability is theorized as the relationship between the *features* of samples—how can models explain *themselves* when inferring? While these methods help ascertain the model's 'mental schema,' there is another tone of explainability that asks *what explains the model acting this way instead of another*, raising questions about the data science process itself. Sample-based explainability understands data science as a causal process that captures data and quantifies it in some way, selects specific data versus others to use in training, and uses an algorithm to construct a computational model which can produce inferences. In this research area, I ask how *intervention in the data science process and the choices made among alternatives lead to a particular model with particular behaviors rather than another*. Besides its impact on regulation, discussed later in the democratic content, this method of explainability can provide quantitative answers to essential questions in active learning and scientific practice, such as: what is the benefit of adding more samples with specific characteristics, what is the impact of experimental noise in the data on specific predictions, how do certain data points provide evidence for particular predictions, and how could new instrumentation or experiments improve model performance? This

[1] Austin Clyde, "Algorithmic Systems Designed to Reduce Polarization Could Hurt Democracy, Not Help It," Tech Policy Press, February 17, 2022, https://techpolicy.press/algorithmic-systems-designed-to-reduce-polarization-could-hurt-democracy-not-help-it/.

can be seen in the context of autonomous discovery initiatives, for example, where active learning and explainability may be interrelated. This work will focus on my work in cancer-drug response prediction, as obtaining new experimental data from animal models or patients is expensive, data is extremely noisy across different models about a real patient response, and explainability is imperative for clinical translation. *Future work:* In this research track, I develop a *data*-centric explainability methods. By reading data science methodology as a causal model for producing data products (fig. 1), great progress can be made on *casual* descriptions of inferences in relationship to the training data. A causal model linking data to model parameters to inferences can also be used to understand the impact of noise in certain data-subsets, potential model improvement with new data, and answer counter-factual questions. Given a model, I will develop a general explainability technique leveraging AI-driven sampling, causal inference, and HPC called Bisection Learning via Automated Data Exploration (BLADE). BLADE will construct a model-specific network of training samples, labels, and inferences to understand how particular inferences are similar and sensitive to the experimental or computational training data (fig. 2). The system will be designed to output specific data linkages in inference mode and be optimized on exascale systems. BLADE will consist of a core data services component, a programmable control module, and a work pool of fine-tuning trainers. Given subsets of the training data make up an extremely large power space, requiring extensive tuning and optimization for high-performance in big data problems, I propose utilizing AI-driven sampling to develop a causal model over the power set of training samples tractable, given the parallels between my prior work in AI-driven adaptive sampling for large molecular dynamics state spaces and the state spaces of parameters which can be sampled under intervention (data-subsetting). This research plan is currently under review as a DOE ASCR call.

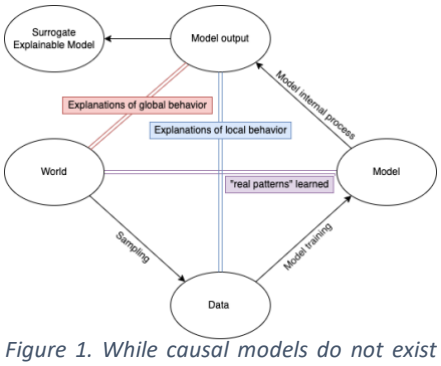

Figure 1. While causal models do not exist generally in data-science problems, a causal model of the model-generation process does.

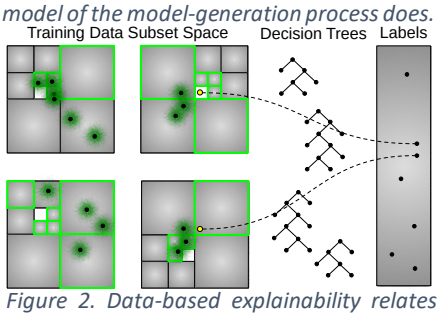

Figure 2. Data-based explainability relates uncertainty in training data to inferences.

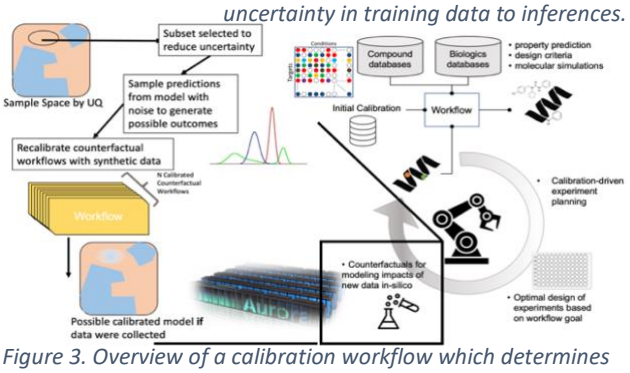

Figure 3. Overview of a calibration workflow which determines the values of new data based on counterfactual model training.

Area 2: Calibration. The COVID-19 pandemic has reified the narrative of declining trust in science. Even in the aftermath of the 2008 crisis, Alan Greenspan's testimony to Congress implied that blame should be apportioned to the lack of calibration. Given moonshot challenges such as developing a computational pandemic preparedness system, such systems must work towards higher calibration standards. Currently, few automated calibration techniques exist for large-scale AI projects, and this divide is only deepening with the scale of models being deployed across diverse industries and science domains. In my work with the National Virtual Biotechnology Laboratory, I scaled computational drug discovery across national and international supercomputing infrastructure; however, this scale revealed several challenges in calibrating such workflows on pre-exascale supercomputers due to lack of methods for hierarchical workflows and disparate data source harmonization. I will focus on developing automatic calibration for large-scale hierarchical AI workflows with existing disparate data sources as well as in cases where experimental data can be acquired (fig. 3). Significant progress has been made with workflows such as IMPECCABLE;[2] however, techniques to auto-calibrate these workflows based on available data, experiments, and/or field observables are lacking. *Future work:* I will develop an HPC framework for the automatic calibration of large workflows that leverage information across the entire campaign and its various scales and optimally discover new data collection(s),

[2] Aymen Al Saadi et al., "IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads," in *50th International Conference on Parallel Processing*, 2021, 1–12.

which will efficiently improve the quality of the model calibration. (1) I aim to decrease the time from pandemic threat detection to effective calibrated exascale workflows with automated computational techniques for bootstrapping calibrated models to emerging threats by transferring calibration information from disparate data sources. I will create a general procedure that produces discrepancy kernels as priors for calibration on unseen tasks by automating the training of Bayesian models on each data set over a harmonized input space on exascale machines. (2) I will decrease the manual time spent on multi-scale workflow calibration by automating the ingestion of existing data and updating calibration through three techniques: transfer learning for uncertainty-calibrated computational data and experimental data sets, workflow-scale campaign managers to calibrate workflows as one differentiable piece. (3) I aim to utilize the calibration data from this research to develop an optimal experimental design protocol for calibrated workflows, prioritizing what data (and what type) should be collected to improve workflow accuracy most effectively. This work has been proposed for a DOE ASCR call on automated calibration.

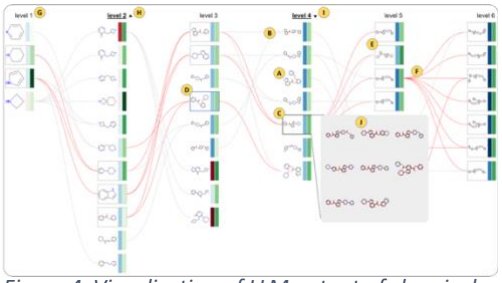

Figure 4. Visualization of LLM output of chemical database

Area 3: Scrutability. How can models be scrutinized as trustworthy? What is suitable due diligence? So far, I have proposed one explainability method to assist this endeavor through grounding inferences in explainable and scrutable data instances. Another challenge for scrutiny is understanding how actors come to interpret the performance, trustworthiness, and "groundedness" of increasingly complex modeling tasks such as large language models (LLMs).[3] As LLMs and other media-generating models such as DALL-E continue to surprise operators with their seemingly oracle-like behavior, I fear those in charge of automated decision systems are not equipped to 'get to know' models and scrutinize their level of groundedness is a present-day reality. Even in AI for science initiatives such as designing autonomous laboratories, LLMs have immense potential to fill current capability gaps. Yet, LLMs' lack of structure or fact grounding presents a significant challenge in science. ***Future work:*** I will research the capacity of LLMs for implicitly and explicitly storing knowledge graphs, treating LLMs as a kind of novel database technology. Conceptually, I believe developing interpretative relationships between models, scientists, and the public can be done through novel visualization inferences and grounding models with structured data (fig. 4). While freeform generation is essential to LLMs' exciting performance, I believe that developing tightly coupled connections to structured databases are essential. A major research challenge is the relationship between training data, their internally stored representation in LLMs, and how they are sampled and transformed. Understanding this relationship will significantly impact misinformation research, LLM regulation, and LLM's application towards automated science. My research has previously explored how language models can be used to traverse a vast graph representing chemical space as a kind of on-the-fly generative database. This LLM-embedded version knowledge graph accurately recalled and generated nodes on the fly over 98% of the time, paving the way for a radical transformation of how databases are implemented. I aim to develop and publicly release a large LLM (>100 billion parameters) in collaboration with Argonne Leadership's Computing Facility. This model is fine-tuned using a graph representation of various scientific databases and knowledge graphs, leveraging query-specific language during training. To quantify the relationship between LLM output and grounded knowledge, I will examine the recall and other characteristics with varying parameter sizes and repetitions in the training data and characterize the recall error. This setup will further allow an understanding of how LLMs produce statements from the training data statements.

Area 4: Democracy. Courts, regulators, and politicians are increasingly relying on computer models and simulations for evidence and reasoning, developing regulation around AI, and grappling with the application of traditional rights in the face of technological artifacts. The ability to engage in communication and collective understanding of policies, form political opinions, and even come to share a unified scientific lifeworld are fundamental to democracy and a basic human interest. How can we afford the same transparency, understanding, and fundamental rights fostered through civic engagement in an increasingly

[3] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models" (arXiv, July 12, 2022), https://doi.org/10.48550/arXiv.2108.07258.

technocratic democracy? In conjugation with developing novel methods for increasing our interpretive flexibility with models through sample-based explainability, calibration, and new methods for scrutiny, I aim to continue scholarship on understanding how interpretations of science and computational follow commitments in law, democracy, and rights. For example, the two modes of explainability I outline coreespond to modes of democratic integration: epistemic and social explainability.[4] Epistemic explainability refers to the traditional X-AI program where the goal is to relate the decision structure of a particular model with a scientific and natural language theory. An explanation might provide a counterfactual ('had expression of gene X increased, the phenotype prediction would have been Y') while a global explanation might present generalized rules to summarize the model ('gene X when expressed with gene Y is phenotype Z''). Social explainability refers to the ways in which different collectives, regulators, and courts come to understand and interpret the system both epistemically and as embedded political, social, and corporate institutions. This distinction is important since social explainability is a pre-condition for democratic discourse in civil society, and it is being realized in law (with a distinction between GDPR Recital 71's call for epistemic explainability and European Union's AI Act art. 13 call for users' ability to "interpret" models). Furthermore, the European Union's AI Liability Directive proposes a presumption of causality when fault has been established with an AI system, and second providing a right to access evidence of a system. By focusing on sample explainability, interpretive flexibility is allowed outside of mere feature attribution. Questions such as the inclusion of some data over other data can be causally understood, allowing open interpretation of those data in social contexts, and quantitively illustrating the impact of those choices. With this charge, I plan on taking on the following projects in the short term: a review of case law focused on the question how is expertise with respect to models used, how intentions are understood in models (as this question arose the recent oral arguments of Merrill v. Milligan), who authorizes what explanations about it, and what might expand the scope of explanation be? Of course, one answer will be provided through technically developing one as above. Second, I will continue research into the kinds of civic epistemologies exercised when it comes to understanding algorithms. Third, I aim to undertake analyzing the parallels between judicial interpretation and interpterion of algorithms to ascertain what commitments follow from ways of knowing algorithms.

Citizens are increasingly faced with 'alien' advanced AI technology. While many AI ethics programs focus on explainability and virtues experts should follow in their practice, few research programs treat the idea with STS reflexivity: *how do these technologies open new means for citizens to participate in world-making, and how can citizens drive the kinds of technological innovation needed in their local contexts?* My research into AI civics is twofold: (1) how do we develop the kinds of public institutions which afford citizens the same access to decision-making that traditional intuitions have in a world of AI? And (2) how do we foster through public education, civic engagement, and university education new skill for an informed citizenry in a technological world. I will articulate AI as an opportunity for empowerment through epistemic justice, where citizens are able to confirm their suspicions and bring new calculability to what oppression is. My work touches human rights law and philosophy, for example, considering the right to the progressive realization of equal access to science and technology.[5] A near-term goal is to take the material from my fall 2022 course—*Artificial Intelligence, Human Rights, and Algorithms*—and writings on the failures of human-in-the-loop ideologies and democracy to develop a manuscript articulating the relationship between AI, human rights, and democracy. I will demonstrate the rapidly changing landscape of scientific practice and policy with AI methods deeply impacts the rule of law, rights, and democracy, provide recommendations for the development of an AI programs that respect these rights, and extend this to human rights philosophy by through a basic interest framework linked to ideas from epistemic human rights.

Conclusion. My future work addresses the question of interpretation in the context of scientific workflows with AI, interpretation of rights with varying accounts of AI, and interpretation of model performance and uses.

[4] Austin Clyde, "Human-in-the-Loop Systems Are No Panacea for AI Accountability," Tech Policy Press, December 1, 2021, https://techpolicy.press/human-in-the-loop-systems-are-no-panacea-for-ai-accountability/.
[5] Austin Clyde, "AI for Science and Global Citizens," *Patterns* 3, no. 2 (February 11, 2022): 100446, https://doi.org/10.1016/j.patter.2022.100446.