esade

Esade Working Paper Nº 278

Regulating AI: lessons from scientific computing

January 2024

Jonathan Douglas Wareham - Ramon Llull University
Angelo Romasanta – Ramon Llull University
Laia Pujol Priego – Ramon Llull University
David Osimo – The Lisbon Council

Do Good. Do Better.

REGULATING AI: LESSONS FROM SCIENTIFIC COMPUTING

ABSTRACT

The integration of artificial intelligence (AI) into various sectors raises critical discussions regarding the regulation of inscrutable AI systems. Drawing parallels from the realm of scientific computing, this essay examines the multifaceted interaction between computation and theory to elucidate lessons for AI regulation and information systems research. Scientific computation has grappled with theory-agnostic approaches that, while inscrutable, have spurred profound advancements and insights. Reflecting on historical and contemporary examples of scientific computing, we propose that the regulatory approach to AI should not be overly fixated on pre-emptive scrutability, but instead must balance ex-ante and ex-post regulatory mechanisms. We argue that strict adherence to explainable AI's idealistic goals may indeed hinder the emergent capabilities of AI applications. We conclude that nuanced regulation informed by the scientific community's uneasy journey with computation can foster responsible, yet innovative, AI deployment.

INTRODUCTION

As artificial intelligence (AI) continues to permeate and impact diverse fields, public apprehension is emerging about a lack of any clear causal logic and transparency in many of its algorithms (Bauer et al., 2023; Lebovitz et al., 2021; Mikalef et al., 2022; Nishant et al., 2023; Ransbotham et al., 2016; Van Den Broek et al., 2021; Zhang et al., 2021). For example, in medicine, computational techniques with no theoretical grounding can outperform extant practices (Kawamleh, 2022; London, 2019; Matulionyte et al., 2022). Likewise, in businesses, if AI can streamline operations or enhance customer service, questions of how it actually works may be considered immaterial. On the other hand, regulatory bodies and civic groups are increasingly posing the question: should scrutability be a prerequisite to the deployment of AI models (Graziani et al., 2023)? Without full transparency into its mechanisms, users and

organizations may not be able to fully control AI, leading to unforeseen consequences, detrimental bias, and outcomes misaligned with public policy and welfare objectives (Bauer et al., 2023; Dwivedi et al., 2023; Meske et al., 2020; Rai et al., 2019).

This paper aims to explore the challenges of AI inscrutability, focusing on an important vanguard of new computational technologies: the scientific community. Scientific research has been a source of many of the world´s most impactful technologies, such as World Wide Web, grid computing, medical imaging, and critical components of most electronic devices (Pujol Priego et al., 2021; Wareham et al., 2022). From its genesis, science has been fixated on theoretical explanations of cause and effect across its various disciplines. However, a closer examination of scientific computation reveals a more intricate relationship between science and explicable, codified theory than commonly assumed. While at its core, scientific computation involves the development of models and simulations to understand natural and social systems; contrary to expectations, the relationship between the scientific enterprise and computing has not always been comfortable or uncontentious. In the early days of computation, the more theory-minded members of some scientific communities viewed computers as glorified slide rules—a rudimentary numerical hack that had no place alongside the aesthetic refinement of formal mathematics (Hammersley & Morton, 1954). It may well be surprising that this equivocal relationship between theory and computation is one that continues today (Mitchell, 2023). And while it is less polemic as it once was, it remains just as nuanced and complex, if not even more so due to recent developments in AI (Benbya et al., 2021; Berente et al., 2021; Mitchell & Krakauer, 2023).

Recent literature has expressed concerns regarding the inscrutability of AI (Berente et al., 2021; Lebovitz et al., 2021; Zhang et al., 2021), the emergence of unexpected properties in deep learning models (Bubeck et al., 2023; Wei et al., 2022), and even the geopolitical and existential threats posed by closed, black-box models (Bostrom, 2017; Bostrom et al., 2020)

‘Explainable AI’ is now the North Star of AI policy debates (European Commission, 2021; National Institute of Standards and Technology, 2023; Phillips et al., 2021; Vilone & Longo, 2021). And while this observation may be unwarranted, one tendency we see in AI policy discussions is a bias toward portraying explainable AI as a dichotomous property: AI logic is either showcased in a well-lit display cabinet, concrete and scrutable, or ensconced in a black box, vaporized into a deep neural net.

Our intention in this essay is not to discount any concerns about AI inscrutability. On the contrary, we consider them serious and, as a response, suggest that *AI regulatory debates can learn a great deal from science’s own struggles with computation, theory, and scrutability*. In science, the relationship between computational techniques and the scientific theories used (or not) for explanation and prediction is far beyond anything dichotomous—it assumes many forms, degrees, objectives, and directions. It is precisely this diversity and nuance that we aim to explore in this essay: how science has employed computers to model, explain, and predict the behaviors of the systems they study, spanning from the 1940s to the present day.

We begin this exploration by describing the assorted history of scientific computation within the post-WWII physics landscape, focusing on the emergence of Monte Carlo simulation (Galison, 2011, 2017). Theoretically, we situate our discussion in the tradition of *instrumentation*, which views computation as a critical instrument that not only detects, measures, or quantifies concepts, but also embodies or defines them (Rheinberger, 2010; van Helden & Hankins, 1994). To substantiate our argument, we examine three examples of scientific computing in high-energy physics, computational chemistry, and structural biology. These examples illustrate three subtly distinct modalities of how computation and theory liaise to increase scientific understanding.

We argue that inscrutability in AI is not just a threat to understanding: it can enable new forms of insight – both directly and indirectly. In doing so, our perspective enriches the broader

discourse in the Information Systems literature which has emphasized the interplay between AI's theoretical grounding (or lack of) and the challenges inherent in the realities of its use (Berente et al., 2021; Lebovitz et al., 2021; Zhang et al., 2021). Recognizing the concerns raised about biases in AI systems (Nishant et al., 2023; Rai et al., 2019), we follow by extracting general insights and normative recommendations for policy debates related to AI development, use, regulation, and ontology. Particularly, we advocate for a responsible, holistic regulatory approach: *AI regulatory policy should seek appropriate balance and coordination between ex-ante and ex-post mechanisms* (OECD, 2021). Through balanced and pragmatic regulatory policy, the potential pitfalls of AI systems can be mitigated without unduly hampering their transformative potential.

COMPUTATION IN SCIENCE: THE EARLY DEBATES

In the vast annals of the history of computing, most early developments in scientific computation were driven by the immediate military needs in cryptography, ballistics, and logistics. The most famous examples being the work of Alan Turing and colleagues at Bletchley Park for cryptography (Cameron 2008), and the Electronic Numerical Integrator and Computer (ENIAC) at the University of Pennsylvania (Smithsonian 2007), designed to perform complex calculations for artillery trajectory tables.

A landmark example was the development of Monte Carlo techniques at the ENIAC computer at the University of Pennsylvania and Los Alamos National Laboratory by John von Neumann, Stanislaw Ulam, Nicholas Metropolis, and others. During post-WWII research on the hydrogen bomb, scientists needed a method to model the diffusion of neutrons through plutonium to calculate the probability of a nuclear fusion reaction (Galison 2011). Such insight was inaccessible experimentally and too complex to be solved with the analytical techniques of the time. The term 'Monte Carlo' is a reference to roulette wheels with random outcomes. These random outcomes can be generated repeatedly, and the resultant distributions can be analyzed

to gain insight into the behaviors of the systems being simulated.

For fusion research, Monte Carlo simulation was developed for pseudo-random number generation to simulate the inherent processes in the hydrodynamical shocks of energy through gas and matter. Neutrons could scatter from a nucleus, be absorbed by a nucleus, or cause a nucleus to undergo fission, emitting an unknown quantity of neutrons. None of these probabilities were calculable via traditional techniques, thus motivating von Neumann and colleagues to refine the Monte Carlo methods for machine computation. The first classified simulations of nuclear fusion were produced using over a million IBM punch cards with the ENIAC computer: The term 'Monte Carlo simulation' was first shared publicly in the *American Statistical Association Journal* in 1949 (Metropolis & Ulam, 1949).

With the growth of computer simulations in science in the 1950s, a new category of research emerged that defied the traditional dichotomy of theorist and experimentalist, with monikers such as ´experimental mathematics' or 'theoretical experiments' (Galison, 2011). It referred to a mode of mechanized analysis for physical systems beyond the reach of both empiricists and theoreticians. Unsurprisingly, this new approach was met with great skepticism and condescension, considered vastly subordinate to the refined world of equations that dominated the Einstein-Dirac apparatus of theoretical physicists (Hammersley & Morton, 1954). It threatened to replace real, theoretical science with low-brow, daily work of computation and numerical approximations (Kowarski, 1972a). Furthermore, legitimate concerns arose regarding replicability and inscrutability, qualities that were otherwise clearly accessible in disciplined mathematical deduction. The computer programs were often too large to be publishable, and the peculiarities of local computing environments limited the portability and replicability of simulations (Galison, 2011). Thus, foreshadowing the concerns of today's AI by some 70 years, these simulations were suspiciously viewed as largely inscrutable, inexplicable, highly sensitive to their deployed architectures, and hence, difficult to repurpose

or redeploy.

However, a deeper metaphysical debate emerged about how they functioned; that is, what was the most appropriate language to describe "scientific truth". The dominant view was that partial differential equations were the most accurate reflection of the Platonic metaphysics governing natural systems. Nevertheless, this view was increasingly challenged by scientists employing new computerized techniques in fields such as chemistry or engineering. These scientists embraced computation as an approach that was equal—and not subordinate to—purely mathematical theorizing (Galison 2017). With the growing adoption and success of Monte Carlo techniques, many scientists began to assert that "stochasticism" (Galison 2011 p.146) was the optimal language to describe the numerous random processes that occur in nature (Kowarski, 1972b). This perspective considers elaborate mathematics as ornamental formalisms that obfuscate a more direct epistemic engagement with the stochastic nature of many natural phenomena (Galison 2011).

So why is this important? As we mentioned, we believe that this brief glimpse into the early debates in scientific computing is useful for framing current deliberations on inscrutable AI. The history of Monte Carlo simulations shows us how many individuals were uncomfortable with the pseudorandom nature of the these back-box simulations. Similar to today's AI, skeptics derided Monte Carlo simulation's lack of any clear causal logic or 'world models' (Mitchell, 2023). Scientists doubted their fidelity to nature and mistrusted their lack of transparency, portability, and replicability. Theorists regarded computers as pedestrian tools—a crude stochastic slide rule—and even discouraged their use whenever possible (Hammersley & Morton, 1954). Yet beyond these more operational polemics surrounding early scientific computation, deeper philosophical deliberations also surfaced.

Even some of Monte Carlo´s staunchest critics eventually recognized that despite (and because of) its atheoretical approach, its simulations could become a new source of insight into

natural problems and systems (Tukey, 1972). In some instances, the deeper value of this unorthodox method was not its conjured predictions, but the subtle hints it could offer for looking at a problem in fundamentally new ways—perspectives that later fed into formal theory. *Shockingly, liberation from extant theory became an asset: Not only did Monte Carlo enable more useful representations and predictions of stochastic processes, but eventually nourished novel theory development*.

As a consequence, computational science slowly rebuilt its status as a crude tool for numerical approximations, gently nourishing it to become a rival, truer language of the world's randomness. With new methods of approximation, representation, and examination, Monte Carlo's atheoretical inscrutability morphed from a threat into a novel embodiment of understanding, an alternative and legitimate language of science with its own ontological value (Galison 2017). These concerns, reverberating from a historical context some 70 years ago, offer an imperfect yet useful analogue that can contextualize and contribute to current AI debates (Birhane et al., 2023).

REGULATION AND INSCRUTABILITY

The current public discourse is ripe with calls for regulation focused on transparency and addressing inscrutability in AI systems (Graziani et al., 2023). The White House just released the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (Executive Order 14110, 2023). In Europe, while a new AI Act is still under negotiation, its content has been drawn up as early as 2021 (European Commission, 2021). Although both frameworks aim to regulate AI, the EU AI Act places greater emphasis on preventing AI's potential surveillance capabilities and safeguarding privacy. For instance, the EU AI Act forbids usage such as social scoring and real-time facial recognition, and it defines strong guardrails for "high-risk" systems, those linked to heavily regulated products and those to be used in critical areas such as biometrics, critical infrastructure, employment,

law enforcement, and migration. The U.S. Executive Order, by contrast, has also a stronger innovation focus, with considerations related to legal liability, intellectual property, and commercializing AI applications (Engler, 2023).

At their core, however, both initiatives seek to place guardrails on the known and unknown risks of AI. Both jurisdictions are concerned that the lack of transparency could lead to a loss of human control: "*human oversight and transparency are an essential element in ensuring that AI systems are in conformity with the relevant legislation*" (Council of the European Union, 2020).

Inscrutability as referred to in the AI literature refers to the challenges of these models being unintelligible to certain audiences (Berente et al., 2021). It is often specified as (1) transparency – whether the code was disclosed by the developers; (2) opacity – whether the algorithm's logic can be accessed; (3) explainability – whether the algorithm's design can enable a description – sometimes causal – of how model predictions are generated.; and (4) interpretability – whether humans can actually make sense of the algorithm and its output in the context of its designed functional purpose (Asatiani et al., 2020, 2021; Berente et al., 2021; Lebovitz et al., 2021; National Institute of Standards and Technology, 2023; Vimalkumar et al., 2021).

However, for many AI applications, these recommendations for *scrutability* might be aspirational – at best. For example, NIST (2023) highlights a possible lack of fidelity or consistency in explanation methodologies, that humans can incorrectly infer a model's operation, or that the model does not operate as expected. Moreover, transparency does not guarantee explainability, particularly if the user lacks sufficient technical skills (NIST, 2023).

Hence, to mitigate the potential risks of AI, The EU AI Act envisages a long set of measures, including: adequate risk assessment and mitigation systems, high quality of the datasets, logging to ensure traceability, detailed documentation, clear and adequate information

to the user, appropriate human oversight, high level of robustness, security and accuracy. These activities are designed to be implemented before the AI solution is publicly deployed as ex-ante measures, aligned with the regulatory culture of product safety. Although the act includes (61) *"post-market monitoring"* and (62) *"reporting of serious incidents and of malfunctioning,"* only 2 of the 85 articles are ex-post, that is, after the deployment of the AI application (European Commission, 2021). The US Executive Order proposes a somewhat more balanced approach between ex ante and ex post measures, although the proposed measures a quite broad, with the intention to define sector specific measures as critical issues emerge (Engler, 2023).

While a proactive, ex-ante emphasis is desirable for policy, many forms of deep neural networks are often excessively complex, making ex-ante, explainable-by-design governance infeasible (Hiriart et al., 2004). As a result, research into explainable AI (XAI) has grown substantially, presenting a variety of methods to understand the behaviors of AI deployments in either testing or operational phases (Fernández-Loría et al., 2022; Kim et al., 2022; Martens & Provost, 2014). Common methods are normally focused on auditing the model's outputs, for example, extracting importance-weights that explain a model's predictions in terms of the influence of specific features as evidence of how the prediction is derived (e.g. LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017)). Other methods include the use of counterfactuals to infer explanations for individual model predictions (Fernández-Loría et al., 2022; Wachter et al., 2017), integrated gradients (Sundararajan et al., 2017), decision trees (Arenas et al., 2022), as well as the development of graphic interfaces to improve human interpretability (Kim et al., 2022). Finally, Mitchell (2023) describes *"probes"* where a simpler neural network is trained to decode the original network's internal activations in response to an input. While all of these methods are valuable, like all diagnostic techniques, they have strengths, weaknesses and sensitivities that render them more-or-less useful depending on the context, and by-and-large,

are just exogenous proxies of a model's internal logic[1]. This is not to suggest that explainable-by-design AI is not a desirable research and policy objective. Rather, simply to acknowledge that the architectures of many AI models make monitoring and compliance of such objectives prohibitively difficult, necessitating balance and coordination between ex-ante and ex-post regulatory mechanisms (OECD, 2021).

THREE TALES OF SCIENTIFIC COMPUTATION

Scientific computation offers a unique perspective on scrutability, explainable AI, and understanding. The larger scientific program for understanding a phenomenon is built on the foundations of instrumentation and theory. Instruments in scientific computation can include hardware (e.g., sensors, detectors), software (e.g., algorithms) and various means of data transmission, storage, analysis, simulation, visualization, and argumentation (Galison, 2008). Scientific instruments are often assumed to agnostically detect, measure, or quantify concepts, but do not in any way embody or define them (van Helden & Hankins, 1994). However, many historians of science consider this perspective naïve (Daston & Galison, 2021; Galison, 2008; García-Sancho et al., 2014). In many forms of scientific research, granular data generated from instruments may lack evidential value until it is integrated and ascribed to some larger theoretical explanation of the natural phenomenon being examined. Hence, theory and instrumentation can become tightly coupled in the machinery of scientific knowledge production[2]. We emphasize this perspective to highlight the potential protagonism that a computational approach can have in embodying scientific meaning in data; that is, the genesis of algorithms that filter, analyze and ascribe evidential value to data are highly interdependent (Pujol Priego & Wareham, forthcoming).

---

[1] It is worth noting that, in some instances, disclosure of the internal logic of a system could infringe on the rights of others by revealing protected trade secrets or violation data privacy data rules (Wachter et al. 2018).

[2] As (Hacking, 1992)) notes, "*Phenomena are not described directly by Newtonian concepts. It is rather certain measurements of the phenomena – generated by a certain class of what might be called 'Newtonian instruments' – that mesh with Newtonian concepts*".

When referring to *theory,* we consider it broadly as the applied conceptual apparatus that explains the functioning of certain aspects of the natural world, describing the regularities inherent in objects and events. Theoretical explanations are a product of many things, including human imagination and mathematical deduction, observation, and experiment. A comprehensive definition of theory is well beyond the scope of this essay. However, for our examples, we focus on cases centered around the prevalent theoretical domains of physics, chemistry, and biology, as the theories are most commonly used in each discipline. What is important for us is that science uses theory as a language of understanding and predicting the behavior of the systems it studies – sometimes referred to as 'world models' in the AI literature (Mitchell, 2023). In pedestrian terms, it is about explaining things with transparent, explanatory or causal logic, that is codifiable, communicable, scrutable, refutable, and useful for others.

In the following, we present three examples of scientific computing that allow us to examine diverse configurations where theory can contribute distinctly to computation (and vice versa). Theory can be foregrounded and significantly define what and how data are generated and interpreted. Conversely, theory can be backgrounded, absent in the algorithmic logic, only called upon to interpret the findings. Finally, the relationship between theory and computation can also be more intertwined, multidirectional, and ambiguous (Karniadakis et al., 2021; Lavin et al., 2021).

Theory in the foreground: CERN

CERN is widely recognized as the world's most powerful infrastructure for high-energy physics research. CERN's application of scientific computing has been driven by an attempt to refute and validate the dominant physics theory known as the Standard Model. To test the predictions, physicists developed ATLAS[3], the largest detector ever constructed for a particle collider, measuring 46 meters in length and 25 meters in diameter. The ATLAS and CMS

[3] About ATLAS: https://atlas.cern/about

detectors[4] were designed to record the high-energy particle collisions of the LHC, which occur at a rate of over a billion interactions per second. On 4 July 2012, ATLAS and CMS announced that they had independently detected a new particle consistent with the predicted attributes of the Higgs boson (Brumfiel, 2012). The Higgs boson is an elementary particle that gives mass to everything in the natural world (Azhari et al., 2020). Discovering the Higgs boson posed a major challenge for particle physics since it cannot be observed directly: it is highly unstable and decays into other particles almost immediately after being produced in high-energy particle collisions[5] (Brumfiel, 2012) rendering it impossible to detect by any straightforward method. First, theoretical physicists predicted the possible decay modes and the probability of each mode occurring. Then, to confirm the existence of Higgs boson, physicists empirically identified these decay products and match their occurrence rates with the theoretical predictions.

Computation plays a crucial role in this process: Complex algorithms were developed to simulate millions of particle collisions and their possible decay pathways. These simulations offered physicists a benchmark to compare data from the LHC's detectors with the theoretical predictions: Computation was critical to transverse the gap between theory and empirical observation. Theory-forward computation tells science what to look for, where and how to look for it, and how to interpret it. Without theory, the massive data quantities generated at the LHC are void of any ontological meaning.

Theory in the Background: Power Storage

Our second illustration comes from the application of generative computational methods in materials synthesis, which have resulted in the discovery of new battery materials

[4] The CMS detector is slightly smaller than ATLAS, and is used to detect similar physics phenomena with different technologies (e.g., magnetic fields) to increase the reliability of the findings. As such, ATLAS and CMS very much work in tandem. https://home.cern/science/experiments/cms

[5] The Higgs boson's exact half-life is unknown, but best estimates place its mean value around $1.56*10^{-22}$. As the Higgs boson has no mass, it travels at the speed of light, c. Multiplying these two numbers produces a distance similar to the diameter of a proton. The diameters of the innermost detectors in the ATLAS and CMS detectors on the LHC are about 30 cm. Hence, the Higgs boson decays well before it reaches any direct detection method.

with improved energy storage capacity, linearity, and longevity (Dwivedi et al., 2023). Traditionally, the discovery of new materials involved trial-and-error approaches requiring time-consuming and resource-intensive efforts. Recent computational applications allow scientists to explore a vast number of potential materials with computational training of existing databases describing known materials and their attributes (Carrete et al., 2014). Computational training unveils patterns in the composition, structure, and attributes of known materials, and based on that training data, it is possible to generate new hypothetical materials with a set of desired properties. The utility of these predictions is the fact that they can be quickly generated and subsequently tested and modified to significantly accelerate the discovery process. Breaking up the process into several steps, scientists first define the properties of the materials that they want (e.g., higher storage capacity, greater stability and linearity, faster charging rates, etc.). Second, scientists compile databases from scientific literature, laboratory measurements, or others sources on existing materials and their known properties and compositions (Jain et al., 2013). Once data is curated and ready, scientists develop different computational models, train them on the data, and extract relationships between a material's composition or structure and its properties (Schmidt et al., 2019). Third, the computational model generates new hypothetical compositions and structures predicted as having the desired properties. In the last step, scientists need to verify the feasibility of the materials (i.e., can they be synthesized at all?) and then experimentally verify if the materials conform to the predicted properties (Finegan et al., 2021).

A landmark example in this area comes from Stanford University, MIT, and Toyota Research Institute, employing computational models to predict the attributes of thousands of new materials for lithium-ion battery cathodes[6]. The research collaboration predicted new types of solid-state battery electrolytes, which promise to revolutionize electric vehicles by offering increased energy storage and safety (Attia et al., 2020). In this example, the role of theory is

[6] Additional information: https://news.stanford.edu/press-releases/2019/03/25/ai-accurately-prl-life-batteries/. And: https://energy.mit.edu/wp-content/uploads/2022/05/The-Future-of-Energy-Storage.pdf

rather implicit: computation does not start with a formal theory of how different elements combine to form a material with certain attributes, but rather, the model learns patterns from data about existing materials, and uses these patterns to predict new materials.[7]

Theory on Demand: AlphaFold

Our third example is drawn from the life sciences and is provided by DeepMind, a subsidiary of Alphabet. In 2018, AlphaFold was introduced to the protein science field as a tool to predict protein structures with an unprecedented degree of accuracy (Jumper et al., 2021). AlphaFold is considered by some to be the most important contribution of AI to date (Toews 2021).

Elucidating the structures of proteins is significant because proteins are the building blocks of life. Proteins fulfill various functions, including nutrient uptake, killing foreign particles, and transporting necessary molecules throughout the body. Fundamentally, proteins are made of a sequence of hundreds to thousands of amino acids, of which there are 22 main types. These component amino acids can then be seen as its alphabet, where different arrangements can produce different structures in 3D space. Such structures are important because they ultimately dictate how the protein functions: The shape enables a protein to interact with high specificity with other biomolecules. With the tight connection between structure and function, scientists have been aiming to map the structure of proteins to deepen the understanding of universal biological systems, especially of human physiology.

The process of empirically determining the protein structure, however, is challenging, typically taking years typically to produce just one validated structure. Experimental techniques such as x-ray crystallography, nuclear magnetic resonance, and cryogenic electron microscopy can only work for some proteins under certain conditions. While computational methods offered

[7] It is important to highlight that, while to a lesser degree, theory still plays an implicit role. For example, the model is trained on a dataset of known materials that have been studied through a long-lasting research tradition rooted in the fundamental theories of physics and chemistry. Hence, one could argue that the model is biased towards such theoretical priors when it learns from the data.

an alternative, it was thought that it was not possible to employ such techniques due to the extremely large number of potential conformations they can take in 3D space. It was thus a huge breakthrough when AlphaFold demonstrated its ability to predict structures with high levels of accuracy; currently with more than 200 million proteins.[8]

When theory is inserted punctually

As a predictor of final protein structures, AlphaFold is successful in predicting proteins that are similar to some 200,000 empirically verified proteins on which it was trained. However, while it succeeds in the metrics of final prediction, AlphaFold offers no insight into how proteins fold in nature. This is largely since the deep learning model's mechanism is theory-free until the absolute final phase, where the predicted structure is checked for violations against known laws of chemistry and physics. Researchers in computational science refer to this as *physics-informed* AI (Karniadakis et al., 2021), where the term 'physics-informed' broadly references any theoretical framework to constrain the model without defining the algorithmic mechanism (Lavin et al., 2021). In fact, a team of researchers attempted to modify AlphaFold by introducing theoretical constraints earlier in the model. They found that including these physical priors (e.g., legitimate chemical bonds) early into AlphaFold's training led to worse performance in the final predictions (Ahdritz et al., 2022).

When computation informs theory

In the same manner that theory can be unidirectionally 'informed' into scientific computing at specific points, AI can reciprocate, where the intermediate and final outcomes of atheoretical computation can enrich theory. This bidirectional transaction between theory and computation is referred to as *physics-infused* AI (Karniadakis et al., 2021; Lavin et al., 2021); once again, where the term 'physics' broadly refers to any relevant theory.

For example, researchers created ExplainableFold, a modification of AlphaFold towards

[8] https://alphafold.ebi.ac.uk/about

counterfactual explanations in protein folding (Tan & Zhang, 2023). Previously, biochemists would attempt to understand the folding mechanism by deleting or substituting small sections in a protein sequence and exploring how these changes affect the protein's overall structure. However, these approaches are slow and difficult to carry out in the lab. Alternatively, ExplainableFold simulates these biochemical experiments virtually, aiding in comprehending the role of different sequences in the overall structure (Romasanta et al., 2023). Similar techniques have also been used to understand protein dynamics; that is, a protein´s movements in 3D space that also determine its function. As seen, AI applied to biology has facilitated theoretical insights that were, to a large degree, unexpected by-products of its main purpose of global structure prediction.

When computation informs empiricism

Mapping the 3-D structure of a protein through experimental methods is a notoriously difficult task. As an example, x-ray crystallography is the most widely used method to empirically determine the structure of many atomic structures, including proteins. A cell biologist attempting to identify the structure and function of a key protein (e.g. a coronavirus) would generate many samples of the proteins and suspend them in a crystal. In the lucky chance that they are successful, this crystal can then be sent to a large synchrotron radiation facility where a powerful x-ray beam is allowed to pass through it. The photons from the x-rays weakly diffract off the electrons in the atoms that form the amino acids chains of the proteins, and the resulting output, if successful, is a 2D scattered pattern of dots called a diffraction pattern. By combining thousands of diffraction patterns captured from various angles, analysts can utilize the dots' locations and intensities to infer the protein's structure. However, this is not straightforward: the diffraction patterns experience notable phase interference (constructive, destructive). Solving this 'phase problem' demands the application of theoretical physics and optics, a traditionally challenging and time-consuming process.

With AlphaFold, this process has been accelerated. Instead of starting from zero, AlphaFold can assist the process by generating basic models of the protein, and, by using this homologue as a model of phase interference, saves researchers substantial time. A back-and-forth workflow that cycles between AlphaFold predictions and empirical validation ensures that the structure is both faster and of higher quality (Read et al., 2023). In this regard, AlphaFold has demonstrated unexpected value by expediting crystallographers' workflows (Romasanta et al., 2023). The value of AlphaFold in this instance is neither final structure predictions nor the replacement of empirical validation; rather, the application of one of its intermediate processes to mitigate a substantial problem that plagued empirical science.

Summary

We have explored three important examples of scientific computation where theory plays a variety of roles. At the LHC, we claim that theory is *foregrounded* and permeates everything: the experimental design, instrumentation, data, and computation are all fully defined by theoretical concepts. Independent of theory, the computational predictions and data have no semantic value. In the Power Storage example, theory is far more discrete and *backgrounded*. Generative AI is trained on known molecular compounds to propose novel molecular structures. Here, theory is absent from the algorithm, backgrounded, and only considered in the subsequent interpretation and validation of the results.

AlphaFold, uses theory more discriminately. Theory is only summoned at the final stage to ensure the predictions are compliant with known chemical laws. Adapting extant concepts in the literature (Karniadakis et al., 2021; Lavin et al., 2021), we call this unidirectional insertion of theory *theory-informed*. Adapting AlphaFold into researchers' workflows through reengineering and adaptations led to several unexpected outcomes. AlphaFold predictions, whether more-or-less accurate, have led to theoretical enrichment beyond the initial purview of single protein structures prediction. Extending the literature (Karniadakis et al., 2021; Lavin et

al., 2021), we call this bi-directional relationship *theory-infused*. Additionally, the intermediate outcomes of the AlphaFold algorithm have accelerated traditional empirical research in methods in X-ray crystallography. Extending the previous concepts, we call this *empirics-infused*. Table 1 summarizes these conclusions.

Table 1. The role of theory in scientific computing

	Theory Foregrounded	**Theory Backgrounded**	**Theory Informed**	**Theory Infused**	**Empirics Infused**
Example	Large Hadron Collider	Power Storage	AlphaFold: Structure prediction	AlphaFold: Region specification	AlphaFold: Phasing and crystalizing
Theoretical locus	Fully based on theory	No theory – generative AI	No theory until final phase	No theory until final phase	No theory until final phase
Locus and role of theory in workflow	Theory predicts type, placement, and frequency of Higgs boson derivatives: what/where/when to look for. How to interpret it.	Interpretation and validation of final results	Imposed in final phase of prediction to ensure validity of the predicted chemical bonds in predicted structures	Theoretical insights into mechanisms of protein folding	Intermediate and final predictions used to refine Xray crystallography phase interactions, identifying homologues, etc.
Directionality & relative position of theory	***Omnipresent*** Theory defines data generation, detection, imaging and computational technologies	***Underlying*** Foundational for interpretation	***Unidirectional: Theory --> AI*** Theory injected into algorithm at specific point	***Bidirectional: Theory --> AI AI -->Theory*** Intermediate insights from predicted structure enrich theories	***Unidirectional: AI --> Empirics*** Intermediate insights from predicted structure combined with extant theory to improve empirical processes

DISCUSSION

Inscrutability and Theory: A More Complex Picture

The main thesis of this essay is that scientific computing's multifaceted, ambiguous, and subtle relationship with theory, while fascinating in its own right, also offers valuable insights for current debates in policy concerning (in)scrutable AI. Undoubtedly, inscrutable AI has undesirable consequences including implicit algorithmic or training data bias (Nishant et al., 2023; Rai et al., 2019), the inability to mitigate outcomes misaligned with public welfare or safety (Mikalef et al., 2022), and increased concerns that relegating human knowledge to AI can cause it to atrophy over time (Bauer et al., 2023; Fügener et al., 2022), further underpinning the existential fears of strong artificial general intelligence (Bostrom, 2017; Bostrom et al.,

2020). However, our brief foray through scientific computing shows that the picture is much more complex.

What we have seen from both the AlphaFold and the Power Storage examples is that a great deal of useful scientific computation can be completed with minimal use of theory, and hence, offers little that is 'explainable'. AlphaFold tells us absolutely nothing about how proteins actually fold in nature[9], yet it has produced a performance unprecedented in structure prediction. Given the likelihood that such outcomes will become more common in science, observers have questioned:

> *"…the implication that a scientific challenge can be considered partially or solved even if human scientists are unable to understand or gain any new knowledge from the solution, thus leaving open the question of whether the problem has been solved."* (Lavin et al. 2021, p. 74).

Hence, our first observation is that ***atheoretical algorithms are not exclusively a threat or liability — liberation from theoretical priors can have value***. There is an understandable (and perhaps necessary) human tendency to mistrust systems that they intuitively do not understand or that diverge from accepted ways of describing phenomena (Lebovitz et al., 2021). The history of Monte Carlo simulations evidences this. Once shunned as an atheoretical, acausal numerical hack, scientists soon learned to appreciate the new stochastic perspectives that more accurately described many natural phenomena as something ontologically valid on its own terms. This naturally leads to the question of whether the 'stochastic slide rule' of Monte Carlo is equivalent to the current 'stochastic parrots' of large language models (Bender et al., 2021). There most certainly are substantial differences. Yet recent experience with AlphaFold shows how its liberation from extant theory not only enables more accurate predictions—but also

[9] Knowledge of actual protein folding is important because many of the protein mutations that cause pathologies happen in the folding process (opposed to genetic mutations).

generates intermediate *theory-infused* and *empirics-infused* outcomes that—unexpectedly—inform both theoretical biology and empirical workflows (Read et al., 2023). In other words, theory-free computation can provide new insights on natural systems, enabling innovative theorizing (Mainzer, 2007). Equally, theory can indirectly gain from computation. The intermediate outputs of computational simulations prove valuable in expediting empirical research that subsequently substantiates and enriches theory development.

The well-known aphorism of (Box, 1976) that 'all models are wrong, some are useful' bodes the possibility that an overreliance on some theories can have restrictive, if not destructive, effects. This has been increasingly acknowledged in medicine:

> *"The long medical preference for radical mastectomy over less aggressive alternatives was driven by the pathophysiological theory that removing as much tissue from the breast as possible would reduce the probability of cancer recurrence… the overreliance on plausible theoretical explanations lead to treatment practices that harmed patients and consumed scarce resources precisely because key causal claims in those theories were false."* (London, 2019).

Apart from medicine, many useful tools and practices have been widely adopted long before there were theories to hypothesize their existence and explain them. Recent examples include superconductivity, neutrino oscillations, and the cosmological microwave background (Krenn et al., 2022). What this suggests is that concerns of ontologically agnostic algorithms in generative AI or LLMs deserve more nuanced consideration beyond skepticism (Galison 2017).

Yet, we also emphasize that discarding theory altogether is also a mistake due to their complementary roles. Hybrid approaches that integrate theory with pattern recognition have made significant breakthroughs, such as in weather prediction (Ebert-Uphoff & Hilburn, 2023). In medicine, virtual organ models combine numerous differential equations to model billions of state variables. With a model of the organ's biophysical properties, the integration of machine

learning models with clinical data have enabled personalized heart simulations (Niederer et al., 2019).

This complex relationship between inscrutability and theory brings us to our second insight: ***We should equally entertain the possibility that unexpected perspectives, methods, and other computational byproducts of atheoretical computing can be both epistemologically useful and ontological legitimate.*** As AI continues to evolve with computational techniques beyond the comprehension of human minds, scientists should embrace a renewed idea of scientific understanding based on qualitative characteristics of the theory that speak to human intuition (De Regt & Dieks, 2005). For example, (Boltzmann, 1964) described gas as a collection of freely moving molecules in a container: As heat increases, so does the motion of the gas molecules, making the gas exert pressure on the container walls. A qualitative sense of temperature and pressure can be gained without any insight into the calculations of statistical mechanics. Yet a growing distance between human intuition and computation is described by (Wang et al., 2023) in their review of AI in the sciences:

> *"The fact that human brains can synthesize high-level explanations, even if imperfect, that can convince other humans offers hope that by modelling phenomena at similarly high levels of abstraction, future AI models will provide interpretable explanations at least as valuable as those offered by human brains."* (Wang et al., 2023 p. 56)

Like computing, science has had a long and complicated relationship with this concept of what 'understanding' a phenomenon actually means (Krenn et al., 2022). For example, physicists can predict the gravitational effects at mezzo and macro levels (e.g., cannon balls and planets) with great accuracy. Nevertheless, they still do not 'understand' gravity on many, many levels[10]. Likewise, the predictive accuracy of quantum mechanics is often heralded as the

[10] Understanding the nature of gravity is considered one of the great unanswered questions of physics, e.g., how it is communicated, its relative weakness compared with other fundamental forces, etc. https://www.livescience.com/34052-unsolved-mysteries-physics.html

pinnacle achievement of modern science. While quantum effects empower the transistors and integrated circuits that enable our ubiquitous handheld digital existence[11], theoretical quantum mechanics is very distant from human perception and hence notoriously difficult to grasp on any intuitive level (Smolin, 2006). Biophysicists can show you the myriad of differential equations that constitute human organ models, but they are likely partial in their scope and depth, and well beyond the understanding of non-experts (Coveney & Highfield, 2023). The point is that even the most successful scientific theories are incomplete: They may predict outcomes with incredible accuracy but have no conceptual notion of process or ontology. Theories may excel in explanatory value and intuitive appeal while offering few refutable predictions. Mathematical formalisms may be aesthetically brilliant yet have little correspondence with terrestrial human perception. All scientific knowledge is provisional, partially useful, and unfinished, and scientists are generally comfortable, if not motivated, by this (Galison 2017).

Hence, this leads us to our third conclusion that ***solving AI inscrutability, while useful, is not a panacea for all challenges of AI explicability***. It is desirable, in that it facilitates insight into how the application functions and derives its outcomes, but it may not directly facilitate an increased 'understanding' of anything on intuitive or scientific levels. Here, the concept of scientific incompleteness can guide the ongoing push from regulators and practitioners for "explainable AI" (Bauer et al., 2023; Pumplun et al., 2023; Vilone & Longo, 2021). Policy premised on a naïve notion of *explicability-equals-understanding* should be updated to acknowledge that all models have strengths and weaknesses, and a model's limitations on inscrutability is just one component.

As such, while we do not contest the value of an ambitious explainable AI policy, our

[11] Transistors rely on quantum effects to function. The MOSFET is the most common transistor used in integrated circuits. It is estimated that 13 sextillion ($13*10^{21}$) have been produced, the most frequently manufactured device in human history (Ledin & Farley, 2022).

examples indicate how scientific computing violates its rather naïve aspirations—frequently—with opaque mechanisms, scant insight, or evidence of how outcomes are derived, thus leaving them partially unintelligible to both scientists and laymen. Nonetheless, they are often successfully reconstructed and repurposed on many levels that prove beneficial for scientific progress. As businesses and society venture further into the realm of AI, any quest for a comprehensive understanding of AI ex-ante may be a missed opportunity. Just like scientific theories, the knowledge we have of AI systems is provisional and constantly evolving. The inconclusiveness of our current AI understanding is both a threat and an opportunity.

INSIGHTS FOR AI REGULATORY POLICY AND IS COMMUNITY

Our preceding reflection may imply that we appeal for a more liberal policy toward AI regulation. This is not the case. What we do argue is that the issue is more subtle than often construed. Particularly, we observe that the ex-ante emphasis of current AI policy, while ideologically desirable, is more likely aspirational than implementable. As a consequence, we argue that ***ex-ante and ex-post mechanisms need to be balanced and coordinated.*** The current landscape of AI suggests a need for realistic mechanisms for ex-post compliance, such as auditing APIs (Wachter et al., 2017) or probes (Mitchell, 2023) that go beyond mere reporting and enable richer regulatory access to AI deployments. This is particularly salient for computational applications with real-world impacts, such as in climate modeling, medical research, and engineering (Heyder et al., 2023). What we want to emphasize is that these problems of inscrutability are not new. The historical canvas of scientific computing is strewn with ontologically-centric or ontologically-indifferent techniques, generating an assorted, imperfect-yet-useful bricolage of evidence around which consensus might eventually form (Galison 2017). What recent, consumerized AI advances have done, however, is to bring such potholed epistemic pragmatism out of the scientific labs and into the smartphones and living rooms of the general public. This obervation suggests that AI policy should be tempered by the

fact that often the users of AI differ substantially, and this naturally should situate discussions on AI policy. Scientific computation is most common in a demographic of users with commensurate expertise to probe, stress, dissect, and reengineer AI, and, frequently, the ability to self-regulate. For the wider populace, regulatory perspectives have different functions and consequences.

Of course, the main reason that AI is now part of the toolbox and lexicon of the general public is the growing investments in AI by large technology corporations such as Alphabet, Microsoft, and Meta (Facebook) as central strategic enablers. As big tech assumes more control over AI's most critical inputs (computing power, large datasets, and highly skilled researchers), it may lead to the prioritization of profit-driven research areas, potentially sidelining more fundamental or socially relevant scientific paths (Ahmed et al., 2023).

Aptly, this Issues and Opinions essay has implications on what can be studied by the information systems community. Historically, it has been observed that numerous technologies originally developed by scientific techniques have subsequently laid the groundwork for information systems in key industrial processes (Romasanta et al., 2021; Wareham et al. 2022). Following this example, we encourage the IS field to rebalance preemptive regulations and equally focus on post-deployment monitoring and control. This extends the discourse around AI regulation to include ongoing governance and adaptability, encouraging research on how organizations can implement robust post-deployment surveillance, auditing, and continuous improvement measures for AI systems. We summarize our insights and implications in Table 2.

Table 2: Updating previous assumptions in AI regulation

Change in assumptions	Implications for policy	Implication for information systems research
Instead of construing AI as mysterious, monolithic black boxes, AI models often offer layers of intermediate processes and outcomes that are made accessible by user communities.	Regulators should avoid excessive efforts to enforce explicability ex-ante as it may constrain the emergent benefits of the model that may only be discovered ex-post.	Researchers can explore the unpredictable but potentially advantageous prospects of inscrutable AI. For instance, research can be done on how organizations and user communities reverse engineer and test numerous attributes for novel insights.
Instead of being static entities, AI models are heavily influenced by	Regulators must consider that AI systems evolve over time due to	Researchers can explore such continuous interrogation of AI

iterative cycles of new data inputs, refining, and self-learning, and continuous interrogation by user communities. Hence Inscrutability is not a static property.	continuous interrogation and adjustment by users. A 'once-and-done' attitude towards regulation is unlikely to suffice: Regulatory enforcement should allow for amendments and updates that reflect evolving AI systems.	through longitudinal research designs across contexts, including organizational standards, risk assessments, and operational protocols that consider the evolutionary nature of AI.
Instead of a one-size-fits-all standard of explainability, AI models require tailorable and context-sensitive explanations.	Regulators should recognize that several factors dictate the required level of explainability, including the sector, application, potential risks, and the user community.	Researchers can explore the varying levels of explainability required in different contexts, as well as on how these needs can be refined and met.
Instead of viewing AI models only in the context of their original intended use, AI models have immense repurposing value, offering alternative modalities valuable for both related and different categories of problems.	Regulators should recognize that atheoretical, inscrutable aspects of AI models are not exclusively threats or liabilities to be mitigated ex-ante: Strict pre-deployment regulations might inhibit the dynamic uses and capabilities of AI. This renders policy premised on ex-ante explicability less feasible and calls for *greater burden on mechanisms for ex-post monitoring and compliance.*	Researchers can study AI repurposing, and move beyond minimizing bias during design and deployment, Research should thus investigate approaches to how such biases are managed ex-post. Researchers can focus on how value is created through repurposing.
Instead of pigeonholing their prime role as a prediction machines, AI models should be viewed as intermediaries in larger processes offering both empirical and theoretical value.	Regulators must recognize the broad context in which AI operates, promoting regulations that cater to both the direct and indirect impacts of AI.	Researchers should revisit the generalization that modern AI is most useful for prediction, while classical techniques are useful for understanding. Attention can be given to designing AI systems that explicitly inform theorizing. Larger questions of how AI reshapes the concept of “understanding” and ontologies in both scientific and business settings are germane.

CONCLUSION

Science, computing, and theory have evolved through an assorted and subtle relationship that has been both contentious and beneficial throughout their history. Understanding something in a scientific sense is far from a monolithic concept. This may render the relationship between computation and scientific explanation as an imperfect simile for policy debates on explainable AI – but therein lies its contribution: AI explicability, as we have argued, is more subtle and complex than often portrayed. More importantly, the atheoretical and inscrutable aspects of many AI deployments are not exclusively liabilities – novel modalities of exploring and simulating the universe, liberated from theoretical priors, offer numerous opportunities that should be acknowledged in AI regulation. This suggests a greater emphasis on ex-post monitoring and compliance than is salient in current regulatory proposals. Policy discussions

on AI regulation can benefit from the nuance and refinement represented throughout the history to the current state of the art of scientific computing to negotiate this delicate equilibrium between ex-ante and ex-post mechanisms.

REFERENCES

Ahdritz, G., Bouatta, N., Kadyan, S. et al. (2022). OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, 2022.11.20.517210. https://doi.org/10.1101/2022.11.20.517210

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, *379*(6635), 884–886. https://doi.org/10.1126/science.ade2420

Arenas, M., Barceló, P., Romero Orth, M., & Subercaseaux, B. (2022). On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems*, *35*, 28695–28707.

Asatiani, A., Malo, P., Nagbøl, P. R., et al. (2020). Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive*, 259–278. https://doi.org/10.17705/2msqe.00037

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., Salovaara, A., (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, *22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Attia, P. M., Grover, A., Jin, N., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, *578*(7795), 397–402. https://doi.org/10.1038/s41586-020-1994-5

Azhari, M., Abarda, A., Ettaki, B., Zerouaoui, J., & Dakkon, M. (2020). Higgs boson discovery using machine learning methods with pyspark. *Procedia Computer Science*, *170*, 1141–1146. https://doi.org/10.1016/j.procs.2020.03.053

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*. https://doi.org/10.1287/isre.2023.1199

Benbya, H., Pachidi, S., Jarvenpaa, S. L., (2021). Special issue editorial: artificial intelligence in organizations: implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281–303. https://doi.org/10.17705/1jais.00662

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3442188.3445922

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Boltzmann, L. (1964). *Lectures on gas theory*.

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135–148. https://doi.org/10.1111/1758-5899.12403

Bostrom, N., Dafoe, A., & Flynn, C. (2020). Public policy and superintelligent AI. In *Ethics of Artificial Intelligence* (pp. 293–326). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0011

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brumfiel, G. (2012). Physicists declare victory in Higgs hunt. *Nature*. https://doi.org/10.1038/nature.2012.10940

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*.

Carrete, J., Li, W., Mingo, N., Wang, S., & Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Physical Review X*, *4*(1). https://doi.org/10.1103/physrevx.4.011019

Council of the European Union. (2020). *Presidency conclusions—The charter of fundamental rights in the context of artificial intelligence and digital change*.

Coveney, P., & Highfield, R. (2023). *Virtual you*. Princeton University Press. https://doi.org/10.1515/9780691223407

Daston, L., & Galison, P. (2021). *Objectivity*. Zone Books. https://doi.org/10.2307/j.ctv1c9hq4d

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137–170.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

Ebert-Uphoff, I., & Hilburn, K. (2023). The outlook for AI weather prediction. *Nature*, *619*(7970), 473–474. https://doi.org/10.1038/d41586-023-02084-9

Engler, A. (2023). *The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment*. https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts, Brussels*.

Executive Order 14110, 75191 (2023).

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining data-driven decisions made by AI systems: The counterfactual approach. *MIS Quarterly*, *45*(3), 1635–1660. https://doi.org/10.25300/misq/2022/16749

Finegan, D. P., Zhu, J., Feng, X., et al. (2021). The application of data-driven methods and physics-based learning for improving battery safety. *Joule*, *5*(2), 316–329. https://doi.org/10.1016/j.joule.2020.11.018

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Galison, P. (2008). Ten problems in history and philosophy of science. *Isis*, *99*(1), 111–124. https://doi.org/10.1086/587536

Galison, P. (2011). Computer simulations and the trading zone. In G. Gramelsberger (Ed.), *From science to computational science* (pp. 118–157). Diaphanes.

Galison, P. (2017) "The pyramid and the ring: A physics indifferent to ontology." In *Research Objects in their Technological Setting,* 10:15-26. London and New York: Routeledge, 2017.

García-Sancho, M., González-Silva, M., Jesús Santesmases, M., & Rheinberger, H.-J. (2014). Shaping biomedical objects across history and philosophy: A conversation with Hans-Jörg Rheinberger. *Dynamis*, *34*(1), 193–209.

Graziani, M., Dutkiewicz, L., Calvaresi, D. et al. (2023). A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, *56*(4), 3473–3504. https://doi.org/10.1007/s10462-022-10256-8

Hacking, I. (1992). The self-vindication of the laboratory sciences. *Science as Practice and Culture*, *30*.

Hammersley, J. M., & Morton, K. W. (1954). Poor man's Monte Carlo. *Journal of the Royal Statistical Society: Series B (Methodological)*, *16*(1), 23–38. https://doi.org/10.1111/j.2517-6161.1954.tb00145.x

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), 101772. https://doi.org/10.1016/j.jsis.2023.101772

Hiriart, Y., Martimort, D., & Pouyet, J. (2004). On the optimal use of ex ante regulation and ex post liability. *Economics Letters*, *84*(2), 231–235. https://doi.org/10.1016/j.econlet.2004.02.007

Jain, A., Ong, S. P., Hautier, G., et al. (2013). Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(1). https://doi.org/10.1063/1.4812323

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Kawamleh, S. (2022). Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics*, *3*(3), 901–916. https://doi.org/10.1007/s43681-022-00212-1

Kim, B., Srinivasan, K., Kong, S. H., Kim, J. H., Shin, C. S., & Ram, S. (2022). ROLEX: A Novel method for interpretable machine learning using robust local explanations. *MIS Quarterly*, *47*(3), 1303–1332. https://doi.org/10.25300/MISQ/2022/17141

Kowarski, L. (1972a). *Computing as a language of physics: Lectures presented at an international seminar course at trieste from 2 to 20 Aug. 1971*. International Centre for Theoretical Physics, International Atomic Energy Agency.

Kowarski, L. (1972b). Computers: Why? *CERN Courier*, 59–61.

Krenn, M., Pollice, R., Guo, S. Y., et al. (2022). On scientific understanding with artificial intelligence. *Nature Reviews. Physics*, *4*(12), 761–769. https://doi.org/10.1038/s42254-022-00518-3

Lavin, A., Krakauer, D., Zenil, H., et al. (2021). *Simulation intelligence: Towards a new generation of scientific methods*. https://doi.org/10.48550/arXiv.2112.03235

Lebovitz, S., Levina, N., & Lifshitz-Assa, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, *45*(3), 1501–1526. https://doi.org/10.25300/MISQ/2021/16564

Ledin, J., & Farley, D. (2022). *Modern computer architecture and organization: Learn x86, ARM, and RISC-V architectures and the design of smartphones, PCs, and cloud servers*. Packt Publishing Ltd.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: *Accuracy versus explainability*. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, *38*(1), 73–99. https://doi.org/10.25300/misq/2014/38.1.04

Matulionyte, R., Nolan, P., Magrabi, F., & Beheshti, A. (2022). Should AI-enabled medical devices be explainable? *International Journal of Law and Information Technology*, *30*(2), 151–180. https://doi.org/10.1093/ijlit/eaac015

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341. https://doi.org/10.1080/01621459.1949.10483310

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mitchell, M. (2023). AI's challenge of understanding the world. *Science*, *382*(6671). https://doi.org/10.1126/science.adm8175

Mitchell, M., Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), https://doi.org/10.1073/pnas.2215907120

National Institute of Standards and Technology. (2023). *AI risk management framework*.

Niederer, S. A., Lumens, J., & Trayanova, N. A. (2019). Computational models in cardiology. *Nature Reviews. Cardiology*, *16*(2), 100–111. https://doi.org/10.1038/s41569-018-0104-y

Nishant, R., Schneckenberg, D., & Ravishankar, M. (2023). The formal rationality of artificial intelligence-based algorithms and the problem of bias. *Journal of Information Technology*, https://doi.org/10.1177/02683962231176842

OECD. (2021). *Ex ante regulation of digital markets*.

Phillips, P. J., Hahn, C. A., Fontana, P. et al. (2021). *Four principles of explainable artificial intelligence*. National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/nist.ir.8312

Pujol Priego, L. Wareham, J. (2023). From bits to atoms: Open source hardware at CERN. *MIS Quarterly*, *47*(2), 639–668. https://doi.org/10.25300/misq/2022/16733

Pujol Priego, L. and J. Wareham (Forthcoming) "Data commoning in life sciences," *MIS Quarterly*. https://doi.org/10.25300/MISQ/2023/17439

Pujol Priego, L., Wareham, J., Romasanta, A., & Rothe, H. (2021). Deep tech: Emerging opportunities in innovation and entrepreneurship. *ICIS 2021 Proceedings*, 3.

Pumplun, L., Peters, F., Gawlitza, J., & Buxmann, P. (2023). Bringing machine learning systems into clinical practice: A design science approach to explainable machine learning-based clinical decision support systems. *Journal of the Association for Information Systems*, *24*(4), 953–979. https://doi.org/10.17705/1jais.00820

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *Management Information Systems Quarterly*, *43*(1), iii–ix.

Ransbotham, S., Fichman, R. G., Gopal, R., & Gupta, A. (2016). Special section introduction—ubiquitous IT and digital vulnerabilities. *Information Systems Research*, *27*(4), 834–847. https://doi.org/10.1287/isre.2016.0683

Read, R. J., Baker, E. N., Bond, et al. (2023). AlphaFold and the future of structural biology. *Acta Crystallographica. Section F, Structural Biology Communications*, *79*(Pt 7), 166–168. https://doi.org/10.1107/S2053230X23004934

Rheinberger, H.-J. (2010). *An epistemology of the concrete*. Duke University Press. https://doi.org/10.2307/j.ctv11qdxmc

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?" In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/2939672.2939778

Romasanta, A. K. S., Wareham, J. D., & Pujol Priego, L. (2023). *Open-source AI: The case of AlphaFold* (ESADE Working Paper).

Romasanta, A. K. S., Wareham, J., Pujol Priego, L., Garcia Tello, P., & Nordberg, M. (2021, July). Risky business: How to capitalize on the success of big science. *Issues in Science and Technology*.

Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. *Npj Computational Materials*, *5*(1). https://doi.org/10.1038/s41524-019-0221-0

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.

Tan, J., & Zhang, Y. (2023). *ExplainableFold: Understanding AlphaFold prediction with explainable AI*. https://doi.org/2301.11765v1

Tukey, J. W. (1972). Data analysis, computation and mathematics. *Quarterly of Applied Mathematics*, *30*(1), 51–65. https://doi.org/10.1090/qam/99740

Van Den Broek, E., Sergeeva, A., & Huysman Vrije, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, *45*(3), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

van Helden, A., & Hankins, T. L. (1994). Introduction: Instruments in the history of science. *Osiris*, *9*, 1–6. https://doi.org/10.1086/368726

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Vimalkumar, M., Gupta, A., Sharma, D., & Dwivedi, Y. K. (2021). Understanding the effect that task complexity has on automation potential and capacity: Implications for algorithmic fairness. *AIS Transactions on Human-Computer Interaction*, 104–129. https://doi.org/10.17705/1thci.00144

Wachter W., Mittelstadt, B.D., & Russell, C. (2018) "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harvard Journal of Law & Technology* 31(2) pp 841-887.

Wang, H., Fu, T., Du, Y., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

Wareham, J., Pujol Priego, L., Romasanta, et al. (2022). Systematizing serendipity for big science infrastructures: The ATTRACT project. *Technovation*. (116), 102374

Wei, J., Tay, Y., Bommasani, R., et al. (2022). *Emergent abilities of large language models*.

Zhang, Z., Yoo, Y., Lyytinen, K., & Lindberg, A. (2021). The unknowability of autonomous tools and the liminal experience of their use. *Information Systems Research*, *32*(4), 1192–1213. https://doi.org/10.1287/isre.2021.1022

esade

ESADE Working Papers Series

Avda. Pedralbes, 60-62
E-08034 Barcelona
Tel.: +34 93 280 61 62

ISSN 2014-8135
Depósito Legal: B-4761-1992