

Digital Humanism, Epistemic Values, and Explanation Toward a Critical Digital Humanism

*Stefania Centrone**
stefania.centrone@tum.de
*Anna-Vanadis Faix***
afaix@alma-mater-eu.eu

ABSTRACT

This article analyzes the concept of digital humanism and asks whether, and in what sense, it can help ensure that digital systems—including algorithms, automated computation, computer-based models, and large-scale data processing—produce not only efficient outcomes but also knowledge that is understandable and explainable. It examines the relation between digital humanism and epistemic values, focusing on explanation as a central requirement of knowledge in digital contexts. Against a narrow conception of explainable AI, the article distinguishes explanation as a technical output from explanation as an epistemic value. On this basis, it develops a conceptual map of explanatory demands in digital contexts and defends a critical form of digital humanism that assesses digital systems not only in terms of predictive success, but also in terms of their capacity to support intelligibility, scrutiny, and revision.

1. Introduction

In recent years, the notion of “digital humanism” has gained considerable prominence in Europe and beyond. It has emerged as a framework for rethinking the relationship between technology, society, and human agency in the digital age, and it is frequently presented as a distinctively European contribution to debates on digital transformation (Nida-Rümelin and Weidenfeld 2022, v). The label is often associated with the idea of a “third way”: an alternative both to the market-driven, consumer-oriented model typically linked to the United States and to state-centred, authoritarian approaches commonly associated with China (loc. Cit.) At the policy level, this orientation is reflected in the European Union’s commitment to a human-centred digital transformation, according to which

* TUM School of Social Sciences and Technology, Germany.

** Eberhard Karls Universität Tübingen & Alma Mater Europaea – ECH, Germany.

technology ought to serve and benefit all people living in the Union. Thus, digital humanism is meant to combine confidence in human reason with an awareness of the limits and risks inherent in digital technologies.

Despite the success of the label and its widespread use as a normative framework intended to guide the development and deployment of digital technologies – its stated aim being the integration of human values into the design of technical systems – significant conceptual difficulties remain. In a 2024 article published in the *Journal of Responsible Technology*, entitled *What Is Digital Humanism? A Conceptual Analysis and an Argument for a More Critical and Political Digital (Post)Humanism*, the Belgian philosopher of technology Mark Coeckelbergh offers a conceptual analysis of digital humanism and draws attention to a certain theoretical underdetermination of the heading. Within this context, he identifies a number of features that appear to characterize digital humanism: (i) the identification, within current digital technologies, of “a tendency to see and treat humans as machines”; (ii) “the idea that humans should keep control over digital technology”; (iii) the claim that “digital technology should be aligned with human goals and values”; and (iv) the requirement of interdisciplinarity, understood as a call for “a more holistic understanding of humanity’s problems” (Coeckelbergh 2024, §2: *6 Components of Digital Humanism*). Without entering into the details of Coeckelbergh’s critique of each of these elements, it can be argued that the problem of securing a precise conceptual determination of the notion of *digital humanism* extends, *mutatis mutandis*, to the type of normativity that digital humanism seeks to endorse.

The normative underdetermination of digital humanism becomes particularly visible when examined from the operational perspective of research funding frameworks. In such contexts, normativity is mobilised across distinct registers – legal compliance, socio-democratic participation, and procedural transparency – without being grounded in a unified normative account.¹ What remains underdetermined is the specific “ought” that governs digital transfor-

¹ See, for instance, the ELSA funding guidelines on digitalization, Big Data, and artificial intelligence in healthcare, where normativity is articulated primarily in *legal* and *regulatory* terms (e.g. informed consent, the right not to know, data ownership and access, transparency understood as traceability or explicability, as well as authorization and liability); or the DFG Guidelines for Project Proposals, where normativity in relation to artificial intelligence appears mainly in the form of requirements of *procedural transparency* and *scientific integrity*.

mation: whether it concerns rights and obligations, democratic procedures, collective goods, or epistemic standards regulating the production and validation of knowledge.

It seems possible to argue that appeals to “human values” remain incomplete unless accompanied by an account of the norms that structure the production, circulation, and assessment of knowledge in digitally mediated environments. In particular, there are compelling reasons to call for a clarification of the epistemic norms that shape digital practices and underlie the functioning of algorithmic systems. Digital systems do not merely act; they classify, predict, infer, and justify. In doing so, they incorporate implicit standards concerning what counts as a good reason, an acceptable inference, or an adequate form of understanding. When these standards remain unarticulated, appeals to “human values” risk remaining largely declarative, while leaving intact epistemic assumptions that may stand in tension with the very ideals of inclusiveness and responsibility that digital humanism seeks to advance. The central claim of this paper is that this gap can be addressed by reconsidering epistemic values not primarily in terms of predictive success or accuracy, but rather in terms of explanation. It poses the question of what counts as reliability, credibility, and trustworthiness in the information circulating on the web. As Coeckelbergh 2024 rightly stresses, despite decades of conceptual work on value-sensitive design, digital technologies are still largely developed in ways that do not substantially involve ethical reflection, not only because of power asymmetries and economic interests, but above all because of uncertainty about how such implementation should be achieved. Moreover, the very idea of embedding values in technological systems raises further questions: “which values, and whose values”? Values are not static entities; they evolve, often in response to technological change itself, which in turn raises the further question of whether one can simply transpose traditional values into new technological contexts without reconsidering their interpretation in light of the transformations brought about by digital infrastructures (Coeckelbergh 2024, §3: *Critical Discussion*).

2. On Epistemic Norms: A Conceptual Framework.

The discussion may usefully begin with a clarification of what is meant by epistemic norms. A useful starting point is to treat them as the (often implicit) standards that make practices of inquiry, belief-formation, and assertion intelligible *as* knowledge-oriented practices. Epistemic norms guide what counts as a

good reason, which kinds of evidence are admissible, which criteria are employed to judge information as reliable or unreliable, and which forms of reasoning are regarded as appropriate within a particular domain. In his influential 1956 essay *Empiricism and the Philosophy of Mind*, Wilfrid Sellars (1912–1989) formulates this idea in a canonical way. He argues that to classify a state as one of knowing is to situate it within the space of reasons – that is, within a network of justifications, inferential relations, but also potential errors and their corrections. As Sellars puts it: “The essential point is that in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says” (Sellars 1956, § 38). Accordingly, to say that “x knows that p” not only implies that x is able to *describe* the situation referred to by “p”, but also that x is able to locate p within the appropriate space of reasons. What thereby comes into view is that knowledge is not merely *descriptive* but essentially *normative*: the move from “x knows that p” to the claim that *p ought to be placed by x within the appropriate space of reasons*. Moreover, Sellars’s point is not simply that justification matters, but that the very concept of knowledge is bound up with answerability to reasons. “x knows that p” presupposes a background grasp of general facts of the form “q is a reason for p”, “p *because* q” and/or “r counts in favor of p,” which in turn does *not* mean that all knowledge must be *inferential* in the narrow sense; rather, it means that its authority depends on a learned capacity to insert epistemic claims within a broader nexus of norms, *including sensitivity to the conditions under which one is entitled to make the claim*. In other words, the epistemic status of a judgment is inseparable from the practices of giving, asking for, and assessing reasons through which we can recognize, sustain, withdraw, or defend it as appropriate.

A further, closely related Sellarsian theme concerns the dynamic character of rationality. Sellars warns that “foundation” metaphors can mislead if they suggest a static structure in which some claims are absolutely secure and the rest are built upon them (Consider the role of primitive concepts and first principles not only in the classical model of science² – from Aristotle and Euclid

² The reference paper on this classical view is Scholz 1930, 27–44.

to Leibniz³ – but also in the axiomatic-deductive model, for example at the beginning of the twentieth century in David Hilbert’s *Foundations of Geometry*.⁴) Instead, empirical knowledge and science are rational because they form “a self-correcting enterprise which can put any claim in jeopardy, though not all at once” (Sellars 1956, § 38). Epistemic norms, in this spirit, include not only standards for *accepting* claims but also norms for *revising* and *repairing* one’s commitments under the pressure of criticism, counterevidence, and recognized error. (Cp. Brandom 2008, 185–186).

The American analytic philosopher Robert Boyce Brandom (1950+) develops this Sellarsian orientation in explicitly inferentialist and pragmatist terms by analyzing the norms implicit in discursive practice (Brandom 1994, Ch. 1, 5–6). On Brandom’s account, what makes an utterance count as an assertion is its role in a social practice structured by giving and asking for reasons (Brandom 1994, Ch. 3, § 1, 142–143). Such practices require that participants undertake and assess distinctive normative statuses (Brandom 1994, 142–143). In particular, Brandom argues that the “game” of reason-giving is unintelligible unless it distinguishes (at least) two dimensions: *commitment* (what one is responsible for, what follows from one’s endorsements) and *entitlement* (what one has a right to claim, what one can justify when challenged) (Brandom 1994, 159–160). Epistemic agents keep track of what others (and themselves) are committed to and whether those commitments are defensible (Brandom 1994, 143). In this way, epistemic norms are not external rules imposed on an otherwise complete activity; they articulate what the activity *is* – the practice of assertion is constituted by norms governing responsibility, challenge, justification, and inferential consequence (Brandom 1994, 171–172). For Brandom, a central task is to make explicit what is implicit in practice: we can bring our inferential commitments under rational control by articulating the rules we are already

³ Cp. Aristotle, *An. Post.* I, 13; Euclid *Elementa*. Also cp. e.g. G. W. Leibniz, *Meditationes de cognitione, veritate et ideis* (1684), AA VI.4A, N. 141; *Non inelegans specimen demonstrandi in abstractis* (1687), AA VI.4A, N. 178. *Nouveaux Essais sur l’entendement humain* (1704), AA VI.6, Book 4, Chapter vii (“Of Maxims or Axioms”), § 11. On the classical model of science from Aristotle to Bernard Bolzano (1781–1848) and Edmund Husserl (1859–1938) and David Hilbert (1862–1943) cp. Centrone 2010, 102 ff. See also Mainzer 1994, 182–195.

⁴ Hilbert 1899. Although the *function* of first principles differs in the classical model of science and in the axiomatic-deductive model (the classical model sets up the axioms under the assumption that they describe an independently given reality, whereas the hypothetico-deductive model posits axioms ad hoc, because they work, and derives their consequences while ensuring that no contradictions are generated), *both* traditions or models begin from *first principles*.

(often tacitly) following, thereby rendering them available for assessment and revision (Brandom 1994, ch. 1, 26-27). Explanation—understood as the practice of laying out inferential nexuses, and incompatibilities—can thus be seen as a norm-governed activity within the space of reasons (Brandom 1994, ch. 3, 172).

The South African philosopher of language and mind John Henry McDowell (1942+) defends the idea that experience itself already belongs to the space of reasons. Normativity does not enter “afterwards” with judgment: it is internal to the way the world is available to the rational subject. This reinforces the idea that rationality, justification, and epistemic responsibility are not added layers, but conditions of intelligibility of knowing. To stand in a cognitive relation to the world means to already be under the normative constraints of reason. In digital environments, epistemic norms operate at several levels. First, they function as standards for assessing information. Which criteria determine whether a piece of content is to be regarded as trustworthy? Reliability, accuracy, relevance, speed, and efficiency are often treated as central indicators, yet each of these terms presupposes a normative orientation. Second, epistemic norms are embedded in procedures of validation and inference. Algorithmic systems incorporate statistical models, ranking mechanisms, and optimisation metrics that determine inclusion and exclusion, prominence and marginalisation. These procedures are not epistemically neutral; they embody assumptions about what counts as relevant evidence and legitimate inference. Third, epistemic norms shape cognitive and social practices. By filtering, ordering, and framing information, digital systems influence the formation of beliefs and the structure of public discourse.

Digital systems, therefore, are not merely technical tools; they structure the conditions of truth and credibility in contemporary digital spaces. They do not determine truth in an objective sense, but generate operational versions of knowledge grounded in data selection, model architecture, and optimisation goals. What is sometimes described as “algorithmic truth” is shaped by design choices. Consequently, digital epistemic norms contribute to shaping cognitive environments, influencing what is taken to be reliable knowledge and conditioning participation in processes of public deliberation.

The central claim of this paper is that a fruitful way of addressing the plurality of senses in which “digital humanism” is invoked is to foreground epistemic norms and to rethink epistemic values not primarily in terms of predictive success or statistical accuracy, but in terms of explanation. Prediction and

accuracy are undoubtedly relevant, yet they are insufficient to capture the normative dimensions of digital practices. What is at stake in many contemporary applications of artificial intelligence is not merely whether outputs are correct, but whether the processes that generate them can be understood, evaluated, and critically scrutinised in epistemically meaningful ways. Explanation, in this sense, is neither reducible to technical transparency nor to post hoc interpretability. It constitutes a normatively charged epistemic practice that structures accountability, agency, and critique within digitally mediated environments.

3. From Explainability to Explanation: Why XAI Is Not Enough for Digital Humanism

It is tempting to treat explainable artificial intelligence (XAI) as a recent response to the problems raised by deep learning and the controversies of the last decades. But the issue of explanations in AI is in fact much older. The demand that artificial systems should be able to answer “why” questions, or render their reasoning intelligible is as old as the ambition to build systems that reason. What is recent is a particular configuration: highly performant statistical models deployed in high-stakes contexts, combined with institutional demands for accountability and contestability, under conditions of proprietary opacity and rapid technological turnover.

In the 1970s–1990s, by contrast, explanation belonged to the grammar of knowledge-based AI. Expert systems ((Mainzer 2020, 11 ff.) were built around explicit knowledge representations – rules, frames, structured domain models – that made the route from premises to conclusions, at least in principle, available for inspection. In medical consultation systems such as MYCIN, explanation facilities were designed to respond to “WHY?” and “HOW?” queries and to provide justifications of diagnostic and treatment recommendations (Shortliffe 1976; Buchanan and Shortliffe 1984). Importantly, explanation was not treated as an addendum to performance; it was intertwined with the acceptability of the system in a professional setting where users were expected to scrutinize reasons.

At the same time, the expert-systems literature makes clear that being “interpretable by construction” is not equivalent to being well explained. Rule traces can be narrow, inflexible, and insensitive to the kind of question being asked; they can also fail to clarify why certain considerations are relevant. This is why the 1980s and early 1990s saw explicit work on explanation as a distinct

design problem. Clancey framed explanation as a bridge between formal inference and human understanding, rather than as a mere printout of rules (Clancey 1983, 216-219; 247-248). Swartout's XPLAIN system aimed to generate justifications grounded in an explicit model of development, and the Explainable Expert Systems framework⁵ (a design paradigm for expert systems that integrates inferential and explanatory capacities as constitutive elements of the system's architecture) pursued explanation as an interactive dialogue supported by richer and more structured knowledge representations⁶ (Swartout 1983; Neches, Swartout, and Moore 1985; Swartout, Paris, and Moore 1991). In a survey that has lost none of its relevance, Moore and Swartout observed that explanation supports not only persuasion but also debugging as well as the discovery of when a system is being pushed beyond the limits of its knowledge (Moore and Swartout 1988).

A second shift occurred with the consolidation of statistical machine learning in the 1990s and 2000s and with the later success of deep learning. Breiman's diagnosis of "two cultures" captures the reorientation: from interpretable data models to algorithmic approaches that treat the data-generating process as unknown and prioritize predictive performance (Breiman 2001). As models moved toward high-dimensional representations, ensembles, kernels, and complex optimization procedures (Bishop 2006; Hastie, Tibshirani, and Friedman 2009), interpretability often ceased to be "built in" and became a separate problem to be solved after training. Deep learning accentuated this tendency: while early technical milestones and later large-scale successes differed in detail, the overall trajectory favored accuracy and scalability over human-legible structure (Hinton, Osindero, and Teh 2006; Bengio 2009; Krizhevsky, Sutskever, and Hinton 2012; LeCun, Bengio, and Hinton 2015).

It is against this background that, by the mid-2010s, XAI emerged as a widely used label for a recognizable field. The label did not merely rename an old concern; it reorganized heterogeneous techniques into a programmatic agenda tailored to modern machine learning. A prominent catalyst was DARPA's Broad

⁵ A design paradigm for expert systems that integrates inferential and explanatory capacities as constitutive elements of the system's architecture.

⁶ not merely lists of rules, but structures that model concepts, relations, and inferential strategies.

Agency Announcement (DARPA BAA16-53 2016),⁷ which defined the aim of developing learning systems that maintain high performance while producing explanations that enable users to understand, appropriately trust, and effectively manage intelligent systems (see also Gunning and Aha 2019). In parallel, regulatory and legal debates intensified the social demand for reasons. The general data protection regulation (GDPR), adopted in 2016, sharpened attention to automated decision-making and to the information duties and safeguards associated with it (Regulation (EU) 2016/679, esp. Art. 22). The much-discussed “right to explanation” became a central focus of a number of recent contributions (Goodman and Flaxman 2017; Wachter, Mittelstadt, and Floridi 2017), and the proposal of counterfactual explanations explicitly linked explanatory practices to the possibility of refutation (Wachter, Mittelstadt, and Russell 2018).

The methodological repertoire of XAI expanded rapidly in response to these pressures. It includes local surrogate explanations that approximate a model’s behavior around a given instance (Ribeiro, Singh, and Guestrin 2016), feature-attribution methods intended to quantify the contributions of input variables (Lundberg and Lee 2017), visualization techniques for deep neural networks (Selvaraju et al. 2017), and counterfactual explanations that describe minimal changes sufficient to obtain a different outcome (Wachter, Mittelstadt, and Russell 2018). At the same time, the field has repeatedly diagnosed its own conceptual instability: what counts as “interpretability” or “explainability”, for whom, and under which success conditions, remains disputed (Doshi-Velez and Kim 2017; Lipton 2018). *Surveys and taxonomies often reveal that the same technique can serve incompatible aims – debugging, persuasion, compliance, scientific insight – depending on context* (Guidotti et al. 2019). Moreover, as Miller has argued by importing insights from social-scientific work on explanation, human expectations of explanation are contrastive, selective, and tied to norms of relevance that are not captured by default technical metrics (Miller 2019).

⁷ The Defense Advanced Research Projects Agency (DARPA) is a U.S. government agency within the Department of Defense that supports research aimed at developing innovative technologies for national security.

Yet the same repertoire also makes visible a structural risk: when explanation is treated as an output, its adequacy is judged by surface features, and explanation slides into persuasion. Explanations can be plausible without being faithful. Saliency maps can highlight visually salient regions while failing to track what the model is actually using; local surrogates can fit well in a small neighborhood while misrepresenting the geometry of the decision boundary; counterfactuals can be actionable yet epistemically misleading if they privilege easily changeable proxies over the underlying determinants of an outcome. This is one reason why some authors have proposed “sanity checks” and robustness tests for explanation methods, asking whether explanations change appropriately when models or labels are randomized (Adebayo et al. 2018).⁸ The deeper point, however, is not that one should merely perfect XAI and move on. It is that explanation cannot be detached from the normative question of what counts as a reason in a given practice. If the point of explanation is to enable verifiability, then stability, fidelity, and the possibility of refutation become part of its epistemic profile.

This is where digital humanism must insist on a conceptual distinction. Much contemporary XAI treats explainability as a functional requirement: the system should output an explanation-like artifact that can be attached to a prediction. This framing transforms a normative question – what type of ought is relevant in a given context, who is owed reasons, what counts as an adequate reason, and under which conditions reasons can be refuted – into an engineering task. It fosters unwarranted conceptual conflation (for example, conflating explanation with interpretability, accuracy with reliability, and transparency with justification). *Transparency* concerns access to internal workings or to the

⁸ In particular, Adebayo et al. (2018) focus on “explanation methods as tools to elucidate important aspects of learned models,” in particular “saliency methods,” a “broad category of visualization and attribution methods aimed at interpreting trained models.” These methods are described as “an increasingly popular class of tools designed to highlight relevant features in an input, typically an image.” The authors show that “*some widely deployed saliency methods are independent of both the data the model was trained on and the model parameters*” (2 ff. Our Emphasis). The moral seems to be that if an explanation method yields similar results even when the model is randomized, or when the labels are randomized, then *the “explanation” is not in fact explaining the model*. Rather, it is tracking something else – for instance, structural features of the input, superficial statistical patterns, or artifacts of the technique itself. The issue, therefore, is not merely to produce explanations, but to ensure that they are genuinely model-dependent and thus epistemically relevant. In other words, some tools of explainability risk seem being epistemically empty.

data and procedures by which a model is produced. It means that the system's functioning should be accessible and open to inspection. *Interpretability* concerns the extent to which the model's structure and parameters can be understood as meaningful by human agents. *Explainability*, in much current usage, concerns the production of post hoc accounts that render an output intelligible in some sense. These are related but not identical, and digital humanism cannot afford to treat them as interchangeable, because each supports different forms of epistemic agency and different modes of critique.

For digital humanism, the relevant contrast is therefore not simply between explainable and non-explainable models, but between *explanation as output* and *explanation as epistemic practice*. Explanation as output treats explanations as products to be generated on demand and evaluated primarily by usability or presentation. Explanation as epistemic practice concerns the situated activities through which agents make sense of claims, judge relevance, and evaluate reasons; it presupposes a question – answer structure, a background of shared assumptions, and a space for follow-up and disagreement. Pragmatic accounts of explanation in philosophy of science already make this point: explanations are answers to why – questions posed by particular audiences, in particular contexts, and they are governed by standards of relevance (van Fraassen 1980; Lipton 1990). On this view, explanation is not a mere “interpretive layer”; it is a normatively loaded practice that enables understanding, justification, and refutation.

Two epistemic values are particularly salient in this transition from output to practice: *intelligibility* and *relevance*. Intelligibility concerns the capacity of an explanation to support understanding by finite and situated human subjects; it is tied to the possibility of asking follow-up questions, identifying scope conditions, and recognizing what the system takes to be evidence. Relevance concerns the selection and articulation of what matters for a given why-question: not everything that correlates with an outcome is explanatory, and not everything explanatory is relevant to the concerns of those affected. Taken together, these considerations show why XAI is insufficient as a foundation for digital humanism. XAI methods can be valuable components of broader explanatory practices, but only if they are situated within norms of intelligibility and relevance, and within institutional contexts where reasons can be requested, examined, and disputed. For digital humanism, the central question is not whether a model can produce an explanation-like output, but whether digital systems can participate in practices of explanation that sustain accountability and critical agency. In

other words, explanation is a condition for limiting trust as much as for producing it, and it must be evaluated as a public, revisable practice.

4. Explanation as an Epistemic Value: What Kind of Explanation? A Conceptual Map

In light of the foregoing considerations, there appear to be sufficient grounds to argue that the normative underdetermination of the concept of digital humanism can be reduced only if we specify which epistemic norms are at stake in digitally mediated practices, and if we stop treating explainability as a merely technical add-on to predictive performance. What is needed is a shift in focus: from explainability as a feature of systems to explanation as an epistemic value and as a public practice. The point is not that explanation replaces other epistemic goods - accuracy, reliability, robustness - but that it reconfigures them. In high-stakes digital environments, *explanation changes what it means to be accurate, reliable, or trustworthy*: outputs are action-guiding, institutionally consequential, and embedded in asymmetric relations of power, and therefore they must be located within a space of reasons (Sellars 1956, 253-329) that makes room for scrutiny, revision and refutation.

Treating explanation as an epistemic value forces a question that the label “XAI” tends to bracket: *explanation of what, for whom, and for which purposes?* In many deployments of machine learning, the *explanandum* is ambiguous. Is the demand for explanation directed at the model (its internal operations), at the target phenomenon (the world-structure that makes the prediction possible), at the specific output (why this individual was classified in this way), or at the broader decision pipeline (why a socio-technical system produced a certain outcome)? These are not interchangeable questions. They require different evidential bases, support different forms of refutation, and raise different issues of responsibility. The central normative lesson is therefore pluralistic: there is no single “the explanation” that can satisfy all epistemic and ethical demands in digital environments, because different audiences require different kinds of reasons and different forms of access.

A defensible way forward seems to be a basic insight from pragmatic accounts of explanation: explanations are answers to why-questions asked by situated agents, and why-questions are typically contrastive and interest-relative (Cp. Lipton 1990, pp. 254-265; Miller 2019, pp. 3-6). “Why P?” often means “Why P rather than Q?”, where Q is a salient alternative in the conversational

and practical context. In algorithmic settings the contrast class is rarely explicit, yet it is normatively decisive. “Why was I denied a loan?” may mean “Why me rather than someone like me?”, “Why denial rather than approval given my profile?”, or “Why this decision rule rather than another that would be more consistent with the institution’s stated commitments?” Making the contrast explicit is already an epistemic achievement: it clarifies what would count as a reason, what would count as a refutation, and what kind of evidence would be needed to evaluate the decision.

To operationalize this pluralism, we propose a conceptual map organized along four dimensions: (i) *the object of explanation* (what is being explained), (ii) *the explanatory relation* (what kind of dependence is involved), (iii) *the audience and purpose* (whom or what the explanation is for), and (iv) *the adequacy conditions* (how the explanation is to be evaluated). The point of the map is not taxonomic elegance but normative guidance. It helps prevent “category mistakes” whereby a technical description of the model’s behaviour (for example, a feature-importance visualization indicating which variables carried weight in the model’s decision) is offered in response to a question that is actually normative (for example, “was this decision justified?”), or whereby a user-facing justification is mistaken for a causal account of the world.

[Ad_i] At least four *explananda* seem to recur in digital contexts: (1) *Output-level explanation* (“Why this output?”). The target is a specific prediction, classification, recommendation, or score. The concern is typically revisability: can an affected subject understand what drove the outcome and meaningfully challenge it? (2) *Model-level explanation* (“How does the model work?”). The target is the learned mapping or decision procedure itself: which representations, features, and learned regularities drive performance? The concern is often validation and debugging, but also scientific insight when models are used as exploratory tools. (3) *Phenomenon-level explanation* (“Why does this happen in the world?”). The target is not the model but the domain: the causal, mechanistic, or structural factors that produce the outcome the model predicts. The concern is understanding and intervention: what can be changed (and how) to change outcomes? (4) *System-level explanation* (“Why did this socio-technical system produce this outcome?”). The target includes data collection and labeling practices, infrastructure, human oversight, institutional goals, and economic incentives. The concern is governance: accountability does not attach only to the trained model but to the pipeline that shapes what counts as evidence and what

counts as success. In the language of “public relevance algorithms”, what is at stake is how procedures for selecting and ranking information become conditions for what can be known and how publics are constituted (Cp. Gillespie 2014, 167-170; 188-191). The distinction matters because many contemporary frustrations around “right to explanation” debates can be redescribed as mismatches between *explananda*. Affected subjects are often offered model-level or output-level explanations when their complaint is system-level (for example, discriminatory data practices), while developers may seek model-level explanations when what is needed is phenomenon-level causal knowledge about which dependencies are stable under intervention.

[Ad_ii] Each object of explanation can be addressed by different explanatory relations. For present purposes, five families are especially relevant: (1) *Causal and interventionist explanations*. These aim to identify difference-makers: factors such that (appropriate) interventions on them would change the outcome. In algorithmic decision-making, causal explanations are crucial when the normative demand is not merely “tell me what you used” but “tell me what would have to be different for the decision to change.” This is why contrastive and counterfactual forms of explanation can be epistemically significant, provided they track stable dependencies rather than merely exploitable correlations (Cp. Woodward 2000, pp. 199-205). (2) *Mechanistic and process explanations*.⁹ These aim to account how an outcome comes about. In digital contexts, however, “how an outcome is produced” can mean two different things: *the mechanism in the target domain* (for example, a social mechanism generating disparities) or the *computational mechanism* (for example, the steps by which a pipeline transforms inputs into a score). In the first case, the explanation concerns the social or material world to which the system is applied. In the second case, the explanation concerns the internal functioning of the technical system. The crucial point is that these two are not necessarily identical. An explanation of the

⁹ Mechanistic explanations are prominent in the philosophy of biology and neuroscience and account for a phenomenon by identifying a system’s components, describing their activities, and showing how their organized interaction produces the phenomenon. A model’s output can be mechanistically explained by referring to its layers, internal transformations, and architectural organization. Process explanations, by contrast, focus on the dynamic sequence of events leading to a result. They emphasize the temporal unfolding of steps from initial conditions to outcome. While mechanistic explanations typically include a process, not all process explanations are mechanistic, since they may describe a sequence without analyzing the underlying structure in detail.

computational functioning is not automatically an explanation of the phenomenon in the world. Mechanistic explanation is often what supports robust understanding because it clarifies not just *that* a factor matters, but *how* it matters (Cp. Machamer, Darden, and Craver 2000, 3-5). (3) *Statistical and predictive explanations*. These explain by locating a case within a pattern, typically probabilistic or model-based. They can be epistemically legitimate, especially when only population-level regularities are available. But they invite a normative risk: statistical relations can be treated as sufficient reasons for action even when they lack causal or justificatory support. In other words, “what usually happens” can illegitimately replace “why”, undermining the possibility of critique and contestation (Cp. Miller 2019, 21-24). (4) *Reason-giving and justificatory explanations*. These are normative explanations: they explain an action or decision by providing reasons that count in favour of it under some standard. In human contexts, justificatory explanations respond to “why did you do that?” by citing intentions, goals, commitments, and norms. In algorithmic contexts, reason-giving explanations are unavoidable whenever a decision is institutionally attributed and must be defended. Yet they cannot be reduced to feature attributions or correlations. A reason-giving explanation must connect the decision to publicly assessable standards (policies, legal norms, professional guidelines) and specify how the decision satisfies or violates them (Cp. Davidson 1963, 685-687). (5) *Genealogical and structural explanations*. These explain by situating outcomes within broader structures: historical trajectories, infrastructures, incentive regimes, and power relations. For many harms associated with digital systems – surveillance, manipulation, epistemic marginalization – genealogical explanations are indispensable because they disclose how “technical” choices can embody and stabilize forms of authority. The point is not technological determinism, but the claim that artifacts and systems can have political qualities in virtue of the forms of organization they require or privilege (Cp. Winner 1980, 123-125; 131-133).

These families are not mutually exclusive; they often need to be combined. But the map helps to make explicit that different explanatory relations answer different epistemic and political needs. When a credit model denies a loan, a counterfactual (“if your income were higher by X...”) may be useful for the individual, but it does not address whether the system is justified, whether the training data encode structural discrimination, or whether the institution’s risk policy is itself acceptable.

[Ad iii] Explanations are not only about content but about function. At least five purposes recur, each associated with a characteristic audience: (1) *understanding* (scientific or practical); (2) *revisability*; (3) *trust and calibration*; (4) *control and intervention*; and (5) *governance and accountability*. Importantly, these purposes may conflict. Explanations optimized for usability may sacrifice fidelity; explanations optimized for persuasion may obscure uncertainty; explanations optimized for debugging may be unintelligible to lay subjects; explanations optimized for contestation may require a level of disclosure that conflicts with proprietary secrecy. Treating explanation as an epistemic value means that these trade-offs are not merely “product decisions” but normative choices that should be made explicit and contestable (Cp. Miller 2019, 6-10).

[Ad_iv] Because explanations are plural and purpose-relative, their evaluation cannot be reduced to a single metric. Still, we can articulate a set of epistemic values that any adequate explanatory practice must negotiate: (1) *fidelity* (truthfulness): the explanation should track the actual dependence that produces the output, rather than merely offering a plausible story; (2) *robustness and stability*: explanations should not be arbitrarily sensitive to small, irrelevant perturbations; (3) *relevance*: explanations should isolate what matters for the question asked, given the contrast class and the normative interests at stake; (4) *intelligibility* and *dialogical openness*: explanations should be comprehensible to their audience and support follow-up questions; (5) *actionability* without manipulation: explanations should support meaningful agency, but not merely enable “gaming”; (6) *reflexivity* about the socio-technical system: explanations should disclose the institutional and infrastructural conditions that shape what counts as evidence and success (Cp. Miller 2019, 10-15; Gillespie 2014, 175-179). This last point brings us back to the relationship between explanation and epistemic norms. Explanations do not merely bring to light reason – consequences relations; they make explicit the standards by which something is counted as a fact worth acting on. To explain is to locate a claim within a space of reasons: to expose it to demands for justification, to counter-arguments, and to revision (Cp. Sellars 1956, 253-329). In algorithmic environments, however, the space of reasons is partially re-engineered: what counts as a reason can be pre-structured by optimization targets, data availability, and platform incentives. Explanation as an epistemic value is therefore not only a cognitive good. It is a way of reclaiming the ability to ask “why?” in a world where many inferences are automated and many classifications are infrastructurally entrenched.

The conceptual map also clarifies why the normative ambitions of digital humanism cannot be satisfied by a narrow transparency agenda. Transparency, interpretability, and post hoc explainability address only fragments of the map, often at the model or output level. A digital-humanist commitment to explanation requires, instead, setting up explanatory practices that are multi-level (output, model, phenomenon, system), plural in explanatory relation (causal, mechanistic, justificatory, structural), and responsive to the needs of different audiences (developers, domain experts, affected subjects, regulators, and the broader public). Only under these conditions can explanation function as an epistemic value in the strong sense: not just as a means to acceptance, but as a normative condition for understanding, contestation, and democratic critique.

5. Toward a Critical Digital Humanism.

If explanation is treated as an epistemic value and not merely as a technical feature (Section 4), then “digital humanism” can be made more determinate by asking what kinds of epistemic standards should govern digital, model-based ways of knowing. In this sense, a “critical” digital humanism can be framed as a programme in philosophy of science: an account of how digital systems function as epistemic instruments – alongside models, simulations, and measurement devices – and of which norms are required for them to contribute to understanding rather than merely to effective prediction. The target of critique is not “technology” as such but a methodological drift: the tendency to treat predictive success as sufficient for epistemic success, and to treat explanation as an optional supplement rather than as a constitutive part of scientific appraisal.

A familiar way to articulate this drift is the distinction between explanatory and predictive modelling. Explanatory modelling aims at representing dependency relations that support understanding and intervention; predictive modelling aims at minimizing prediction error, often without commitment to the truth of the represented structure (Cp. Shmueli 2010, 289–310). This distinction structures what counts as a relevant variable, what counts as a successful fit, how uncertainty is handled, and what kind of generalization is sought. In the digital sciences and in machine-learning deployments, the two aims are often conflated: high predictive accuracy is taken to license explanatory conclusions, while explanatory narratives are sometimes offered as if they were validated simply by good out-of-sample performance. A critical digital humanism, in the

present sense, begins by insisting that prediction and explanation answer different questions, and that the norms for evaluating them differ accordingly. This insistence is not tantamount with a demand for “transparent” models at any cost. Philosophy of science has long recognized that complex and partially opaque instruments can nevertheless support knowledge – consider the role of computer simulation in contemporary science. Simulations are frequently used as surrogates for experiments, yet the epistemic warrant they provide depends on a network of practices – verification, validation, calibration, and sensitivity analysis – rather than on a single decisive test (Cp. Winsberg 2003, 105–125). The methodological lesson seems appears to be of a general character: where epistemic opacity is unavoidable, the relevant question is not whether the inner workings are fully surveyable, but which practices and standards can stabilize the inferences drawn from the artifact.

Machine-learning systems can be interpreted as a new chapter of this broader story of model-based science. Their opacity is often discussed as if it were exceptional, but many scientific instruments are opaque in the sense that no single agent can cognitively “grasp” all the transformations that connect inputs to outputs. What is distinctive in contemporary machine learning is rather the ease with which a successful predictive pipeline is reified into an apparently autonomous epistemic authority. Here the conceptual map from Section 4 matters: it helps to separate (i) local, output-directed explanations, (ii) model-level accounts of representational and statistical structure, and (iii) phenomenon-level explanations that target causal and mechanistic dependencies in the world. A critical stance consists in preventing illegitimate slides across these levels—for example, from the empirical fact that a classifier performs well to the epistemic claim that it has uncovered the relevant causal structure. This is precisely the kind of inference that philosophy of science treats with caution: models can be empirically adequate without being explanatory, and they can support correct predictions while being systematically misleading about the dependencies that matter for intervention.

A second pillar of a critical digital humanism concerns the status of data. Data are not “given”; they are produced, selected, cleaned, and rendered portable through infrastructural and representational work. In data-intensive research, the evidential role of a dataset depends on provenance, measurement conventions, and the fit between what is measured and what is meant to be inferred. In a paper written in 2016 Sabina Leonelli has emphasized that data travel only by being packaged with assumptions, standards, and metadata that

enable re-use across contexts; when that informational envelope is lost, apparent objectivity can mask deep context-dependence (Cp. Leonelli 2016). This is directly relevant to machine learning pipelines, where training data are often treated as interchangeable tokens of “ground truth”. A philosophy-of-science perspective reframes this as a problem about evidential relevance: which features are measuring what, under which conventions, and with which domain shifts and confounders? On this view, “explanation” cannot be restricted to post hoc accounts of model output; it must include explanations of data generation and of the inferential bridges that connect measurements to claims.

Third, the critical stance requires explicit norms for robustness. Robustness is not simply “accuracy on a test set”; it concerns stability of inferences under reasonable variations in modelling assumptions, data preprocessing, and background conditions. In interventionist accounts of explanation, explanatory relations are expected to display invariance: they should support counterfactual reasoning that remains stable under relevant changes (Cp. Woodward 2000, 197–254). In mechanistic accounts, explanation requires an articulation of organized dependence that can withstand perturbations without collapsing into a merely statistical summary (Cp. Machamer, Darden, and Craver 2000, 1–25). These ideas suggest a practical criterion for digital-humanist critique: when an “explanation” changes drastically under small, irrelevant perturbations of the input, of the training data, or of the model class, it fails to support understanding even if it supports user satisfaction. Robustness tests and sensitivity analyses are therefore not optional engineering extras; they are part of the epistemic profile of explanation.

A related point concerns uncertainty and scope conditions. In simulation-based sciences, it is standard to distinguish between internal checks (verification of code and numerical implementation) and external checks (validation against relevant phenomena), while recognizing that neither step yields an absolute guarantee of truth (Cp. Oreskes, Shrader-Frechette, and Belitz 1994, 641–646). For machine learning models, the analogue is the gap between benchmark performance and performance under domain shift: cross-validation on a fixed dataset can conceal how fragile a model is when measurement practices, populations, or background conditions change. A critical digital humanism therefore treats “generalization” – for example, from previously encountered cases to future ones – not as an automatic consequence of scale, but as a defeasible claim to out-of-sample validity, requiring explicit stress testing and systematic uncertainty reporting.

A fourth pillar concerns inductive risk and the normative structure of methodological choices. Many modelling decisions – choice of loss function, threshold setting, class weighting, or the tolerated rates of false positives and false negatives – encode preferences about which errors are more costly. In a paper written in 2000 entitled *Inductive Risk and Values in Science*, Heather Douglas famously argued that such choices cannot be made value-free whenever the consequences of error are asymmetric; values legitimately enter through the management of inductive risk (Cp. Douglas 2000, 559–579). In a philosophy-of-science framing, this point can be detached from any specific substantive agenda: it is a claim about epistemic responsibility. Digital systems operationalize error trade-offs at scale, often by default or by inherited metrics. A critical digital humanism therefore requires that these trade-offs be made explicit, justified, and revised in the light of the relevant domain-specific knowledge. Otherwise, the system's performance numbers function as a screen that hides the normative commitments implicit in "optimization".

Finally, critical digital humanism must be understood as a proposal about epistemic infrastructure. Scientific objectivity is not secured by individual cognitive access to a model's internals; it is sustained by practices that enable criticism, replication, and error detection. In her *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Helen Longino develops an account of objectivity as a product of "transformative criticism," according to which objectivity is achieved when claims are exposed to structured, competent, and responsive critical interaction (Longino 1990). In machine learning contexts, the analogous infrastructure includes documentation practices that make datasets and models assessable: for instance, "datasheets" that record how datasets were collected, processed, and intended to be used (Cp. Gebru et al. 2018), and "model cards" that disclose performance characteristics, intended contexts of use, and known limitations (Cp. Mitchell et al. 2019, 220–229). These devices can be read as contemporary counterparts of the "materials and methods" sections of scientific papers: they do not themselves provide explanations of phenomena, but they sustain the conditions under which explanatory and evidential claims can be responsibly evaluated.

Taken together, these considerations yield a philosophy-of-science interpretation of critical digital humanism. "Critical" does not mean suspicious by default; it means normatively explicit about what counts as evidence, what counts as explanation, and what counts as warranted generalization in digital settings. It also means resisting the reduction of epistemic success to predictive

success. Digital systems can and do support genuine understanding – but only when embedded in practices that track provenance, test robustness, manage inductive risk, and expose claims to criticism. In this sense, explanation as an epistemic value is not the end-point of the argument but the bridge: it connects digital humanism’s appeal to “human values” with a more precise account of the epistemic norms through which digital systems can contribute to knowledge. This, in turn, provides the basis for articulating how a critical digital humanism can orient the assessment of digitally mediated epistemic practices without lapsing into either uncritical optimism or blanket skepticism.

6. Conclusion

Digital humanism is often presented as providing a moral orientation for the digital age; however, its practical implications for design, research priorities, and evaluation may not always be fully articulated. A promising way of lending greater determinacy to this framework lies in making explicit the epistemic norms that digital systems presuppose and reshape: the standards that regulate what counts as acceptable evidence, a sound inference, and a justified claim in data-driven environments. Our central proposal was to treat explanation not as an optional add-on to predictive performance, but as an epistemic value and a public practice. “Explainability” is frequently reduced to producing an explanation-like output attached to a prediction. By contrast, explanation, understood as giving and asking for reasons, enables scrutiny and revision. This also clarifies why there is rarely a single explanation that can satisfy all legitimate demands. Different contexts require different objects of explanation (an individual output, a model, a phenomenon in the world, or a whole socio-technical pipeline), different explanatory relations (causal, mechanistic, statistical, justificatory, or structural), different audiences and purposes (understanding, contestation, calibration of trust, intervention, governance), and different criteria of adequacy (fidelity, robustness, relevance, intelligibility, and actionability). Making these distinctions explicit helps prevent a common failure mode: answering a normative question with a merely technical artifact, or presenting a convenient justification as if it were a genuine account of how and why an outcome occurred.

On this basis, we outlined a “critical digital humanism” in a philosophy-of-science sense. The core idea is straightforward: predictive success is not the same as epistemic success. Digital systems can function as epistemic instruments, but only when embedded in practices that make their outputs assessable

as knowledge claims rather than treated as opaque verdicts. This involves, at minimum, careful attention to how data are produced and travel across contexts, robustness and sensitivity testing, explicit reporting of uncertainty and scope conditions, and transparency about error trade-offs where the costs of mistakes are uneven. It also requires an epistemic infrastructure that supports criticism: documentation that makes datasets and models auditable, and institutional settings in which reasons can be requested, inspected, and challenged.

The conclusion is therefore practical as much as conceptual. If digital humanism is to be more than a slogan, it should be evaluated by whether digital systems and institutions sustain the conditions for good explanations: truthfulness, relevance, clarity, and the possibility of revision. A human-centered digital transformation cannot be secured only by aligning systems with abstract “values”. It also requires that model-based claims remain inside the space of reasons: exposed to questions, supported by evidence that can be examined, and open to correction when they fail.

REFERENCES

- Adebayo, Julius, Justin Gilmer, Ian Goodfellow, Moritz Hardt, Been Kim, and Jascha Sohl-Dickstein. 2018. “Sanity Checks for Saliency Maps.” *Advances in Neural Information Processing Systems* 31.
- Aristotle (ca. 350 BCE). 1964. *Analytica Posteriora*. Edited by W. D. Ross. Oxford: Clarendon Press.
- Bengio, Yoshua. 2009. “Learning Deep Architectures for AI.” *Foundations and Trends in Machine Learning* 2(1): 1–127.
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Brandom, Robert B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.
- Brandom, Robert B. 2008. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: Oxford University Press.
- Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16(3): 199–231.
- Buchanan, Bruce G., and Edward H. Shortliffe (eds.). 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.
- Clancey, William J. 1983. The Epistemology of a Rule-Based Expert System: A Framework for Explanation. *Artificial Intelligence*, 20(3), 215–251.
- Coeckelbergh, Mark. 2024. What Is Digital Humanism? A Conceptual Analysis and an Argument for a More Critical and Political Digital (Post)Humanism. *Journal of Responsible Technology*, 17. <https://doi.org/10.1016/j.jrt.2023.100073>.

- DARPA. 2016. Explainable Artificial Intelligence (XAI). Broad Agency Announcement DARPA-BAA-16-53.
- Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60 (23): 685–700. <https://doi.org/10.2307/2023177>.
- Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." arXiv:1702.08608.
- Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–579. <https://doi.org/10.1086/392855>.
- Euclid (ca. 300 BCE) (1883–1888). *Elementa*. Edited by J. L. Heiberg. Leipzig: Teubner.
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. "Datasheets for Datasets." arXiv:1803.09010.
- Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: The MIT Press. <https://doi.org/10.7551/mitpress/9780262525374.003.0009>.
- Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'." *AI Magazine* 38(3): 50–57.
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51(5): Article 93.
- Gunning, David, and David W. Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." *AI Magazine* 40(2): 44–58.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.
- Hilbert, David. 1899. *Grundlagen der Geometrie*. Leipzig: Teubner.
- Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18(7): 1527–1554.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521(7553): 436–444.
- Leibniz, Gottfried Wilhelm. 1684. *Meditationes de cognitione, veritate et ideis*. In: *Sämtliche Schriften und Briefe (Akademie-Ausgabe)*, Reihe VI, Bd. 4A, N. 141. Berlin: Akademie-Verlag, 1999.
- Leibniz, Gottfried Wilhelm. 1687. *Non inelegans specimen demonstrandi in abstractis*. In: *Sämtliche Schriften und Briefe (Akademie-Ausgabe)*, Reihe VI, Bd. 4A, N. 178. Berlin: Akademie-Verlag, 1999.

- Leibniz, Gottfried Wilhelm. 1704. *Nouveaux Essais sur l'entendement humain*. In: *Sämtliche Schriften und Briefe (Akademie-Ausgabe)*, Reihe VI, Bd. 6. Berlin: Akademie-Verlag, 1962.
- Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.
- Lipton, Peter. 1990. "Contrastive Explanation." *Royal Institute of Philosophy Supplement* 27 (March): 247-266. <https://doi.org/10.1017/S1358246100005130>.
- Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *Communications of the ACM* 61(10): 36–43.
- Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.
- Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.
- Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67 (1): 1-25. <https://doi.org/10.1086/392759>.
- Mainzer, Klaus 1994. *Grundlagen der künstlichen Intelligenz*. In *Computer – Neue Flügel des Geistes? Die Evolution computergestützter Technik, Wissenschaft, Kultur und Philosophie*, 102–185. Berlin/Boston: De Gruyter.
- Mainzer, Klaus. 2020. *Artificial Intelligence – When Do Machines Take Over?* Berlin/Heidelberg: Springer (Technik im Fokus). <https://doi.org/10.1007/978-3-662-59717-0>.
- McDowell, John. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.
- Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1-38. <https://doi.org/10.1016/j.artint.2018.07.007>.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 220–229. New York: ACM. <https://doi.org/10.1145/3287560.3287596>.
- Moore, Johanna D., and William R. Swartout. 1988. *Explanation in Expert Systems: A Survey*. ISI Research Report ISI/RR-88-2. Information Sciences Institute, University of Southern California.
- Neches, Robert, William R. Swartout, and Johanna D. Moore. 1985. "Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development." *IEEE Transactions on Software Engineering* 11(11): 1337–1351.
- Nida-Rümelin, Julian and Nathalie Weidenfeld. 2022. *Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*. Cham: Springer. <https://doi.org/10.1007/978-3-031-12482-2>.

- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union L119, 4 May 2016.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16), 1135–1144.
- Scholz, Heinrich. 1930. *Die Axiomatik der Alten*. In: *Mathesis Universalis*. Darmstadt: Wissenschaftliche Buchgesellschaft, 1969,.
- Sellars, Wilfrid. 1956. Empiricism and the Philosophy of Mind. In Herbert Feigl and Michael Scriven (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. I: *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, 253–329. Minneapolis: University of Minnesota Press.
- Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.
- Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3): 289–310. <https://doi.org/10.1214/10-STS330>.
- Shortliffe, Edward H. 1976. *Computer-Based Medical Consultations: MYCIN*. New York: Elsevier.
- Swartout, William R. 1983. “XPLAIN: A System for Creating and Explaining Expert Consulting Systems.” *Artificial Intelligence* 21(3): 285–325.
- Swartout, William R., Cécile Paris, and Johanna D. Moore. 1991. “Explanations in Knowledge Systems: Design for Explainable Expert Systems.” *IEEE Expert* 6(3): 58–64.
- van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford University Press.
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.” *Harvard Journal of Law & Technology* 31(2): 841–887.
- Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.” *International Data Privacy Law* 7(2): 76–99.
- Winner, Langdon. 1980. “Do Artifacts Have Politics?” *Daedalus* 109 (1): 121–136. Stable URL: <http://www.jstor.org/stable/20024652>.
- Winsberg, Eric. 2003. “Simulated Experiments: Methodology for a Virtual World.” *Philosophy of Science* 70 (1): 105–125. <https://doi.org/10.1086/367872>.
- Woodward, James. 2000. “Explanation and Invariance in the Special Sciences.” *The British Journal for the Philosophy of Science* 51 (2): 197–254. <https://doi.org/10.1093/bjps/51.2.197>.