

Law of Large Numbers: Accuracy as a Statistical Measure for AI Compliance and Competition

Rabanus Derr^{1, 2}, Alina Wernick^{1, 3}, Robert C. Williamson^{1, 2}

¹ University of Tübingen

² Tübingen AI Center

³ CZS Institute for AI and Law

rabanus.derr@uni-tuebingen.de, alina.wernick@uni-tuebingen.de, bob.williamson@uni-tuebingen.de

Abstract

The machine learning community progresses (in part) by improving the “accuracy” of its systems. The EU AI Act explicitly refers to “accuracy” as part of its compliance measures for high-risk AI systems. Are we talking about the same thing? This work presents “accuracy” as a case-study for differing requirements of social worlds, the technological machine learning community and the legal community. While competition on accuracy contributes to technological development, machine learning scholars simultaneously recognize accuracy’s shortcomings regarding the usefulness and effectiveness of machine learning systems. The legal counterpart embraces the vagueness of “accuracy,” leaving interpretative flexibility for technological and societal changes. At the same time, accuracy is a core element of compliance within the EU AI Act. We elaborate on five main tensions, (a) nature of accuracy, (b) notion of performance, (c) scope of validity, (d) ends, and (e) statisticalness, to show that the two communities project disparate, and sometimes contradictory, expectations on accuracy. Both legal and technical communities lack precise understanding of “accuracy” beyond the contextual boundaries of their community. The resulting frictions, e.g., based on the empirical or normative understanding of accuracy, are symptoms of an unresolved (and unresolvable) debate on what accuracy is. We constructively use the frictions to recommend baselines and interventional studies in standardization, and demand for tools to extend the validity of accuracy measurements.

1 How to Read This Paper?

This paper is structured unlike many others. There are two introductions. Plus, there are two threads in the main section of this paper (Section 3 L and Section 3 E). Are you a lawyer or rather from the social sciences? Then, take the left introduction on the following page. This introduction is marked with an “L” for “law”. Are you a machine learning (ML) researcher or engineer? The right introduction on the following page is for you. You’ll be guided by an “E” for engineer.¹

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Such experimental approaches in scientific literature might be less common in machine learning. In legal research, there exist great examples of breaking the boundaries of standard scientific approaches, documents and formats, e.g., (Fuller 1949) and (Kang and Buchner 2004).

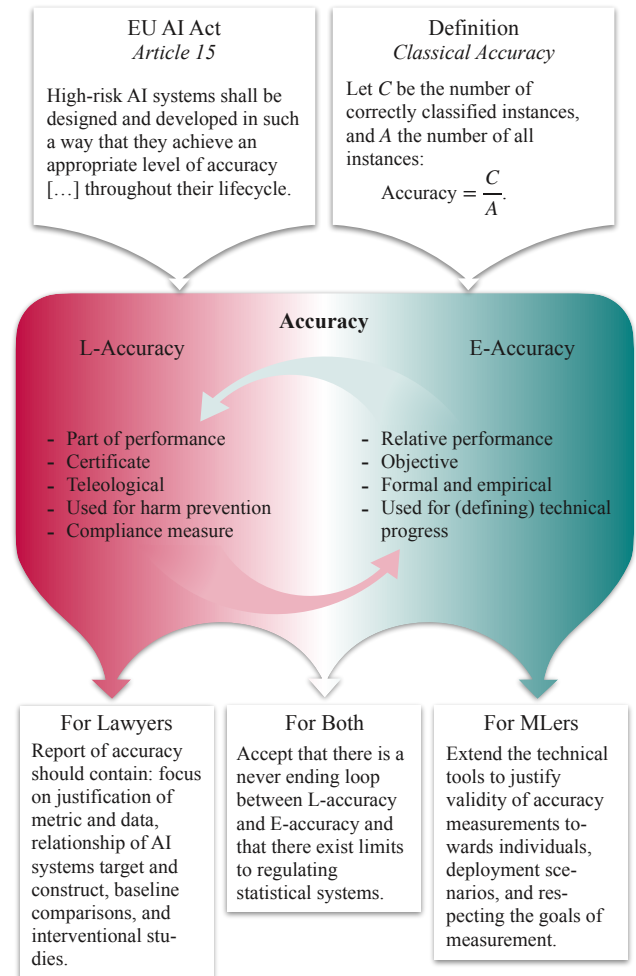


Figure 1: Graphical Summary of the Paper.

1.1 L Introduction for Lawyers

In 2024, the European Parliament passed the EU Artificial Intelligence Act (AI Act, AIA). Enacted in response to a growing concern that AI ethics and self-regulation were insufficient for protecting against risks posed by AI (Smuha and Yeung 2025; European Commission 2020), it regulates the provision and deployment of AI systems in the EU. As a piece of EU technology law, the AI Act follows the blueprint of product safety regulation and aims at the harmonization of the internal market with the aim of supporting the adoption of trustworthy AI. Whereas the earlier product safety regulations sought to protect against risks to health and safety, the AI Act also addresses risks to fundamental rights (Almada and Petit 2025). The drafting of compliance standards for the EU AI act proved challenging. As a result, application of its obligations for high-risk systems, including accuracy, was postponed to 2027 and 2028 (EU 2026).

Besides legal concepts, the AI Act also features terminology that originates from technical community (Koivisto, Koulu, and Larsson 2024). Some of them, such as “explainability” (Kaminski 2021) are explicitly normative. However, others such as “robustness” appear more technical (Nolte, Rateike, and Finck 2025). One such central technical concept is “accuracy”. Accuracy is part of the pivotal compliance obligations for high-risk AI systems (AIA Art. 15), i.e., AI application contexts particularly sensitive to risks to health, safety and fundamental rights, such as education and healthcare (AIA Art. 6; Annex III). For example, the provider of a high-risk AI system must ensure that the system performs with a consistent level of accuracy throughout the life-cycle (AIA Art. 15 (1); rec 73).

Furthermore, accuracy is a cornerstone of current advancements in machine learning (Donoho 2024; Hardt 2025). For engineers, accuracy is a (numerical) attribute of an *algorithm* (a mechanical procedure) or a *model* (essentially a complex mathematical structure) that is designed to make predictions, recommendations, or comparable outputs. Its “accuracy” is a mathematical formula (one of many possible choices) that evaluates the performance on particular test cases.

The engineer strives to make an accurate algorithm in the same way she strives to make a fast car.

Given its centrality to both the legal and the technical community, one would expect “accuracy” to be defined within the EU AI act. It’s *not*. Hence, **what is accuracy?**

In this paper, we argue that a naive deferral of “what is accuracy” to the machine learning community or any self-contained doctrinal approach to the question is doomed to fail. The technical and legal community project different expectations on accuracy, use it in different ways, and follow different aims, cf. (Nolte, Rateike, and Finck 2025). However, neither can claim to converge to a true single notion. Rather, the friction between the expectations reveal the constructive potential of “accuracy”. In other words, we use “accuracy” as a case study for identifying translation gaps, misunderstandings and needs at the intersection of the technological world of machine learning and the social world of law.

1.2 E Introduction for ML Scholars

AI systems have left the laboratory and entered society. They have an impact on economic growth (Aghion, Jones, and Jones 2017). They transform jobs (Acemoglu et al. 2022). They raise (new) concerns about social fairness and the perpetuation of bias against marginalized groups (Barocas and Selbst 2016; Barocas, Hardt, and Narayanan 2023; Hu 2026). They prompt questions about legal accountability and the right to explanation in automated decision-making (Wachter, Mittelstadt, and Russell 2017), and they have led to dedicated regulatory frameworks (European Parliament and Council 2024). In these developments the machine learning community takes on a new relevance. It is not only the technology itself that is adopted and deployed, but also its terms, concepts, and assumptions.

A key concept and value(!?) (Birhane et al. 2022) of machine learning is “accuracy”. Accuracy underlies many recent advances via the wholesale adoption of benchmarking as the style of argument in machine learning (Donoho 2024; Hardt 2025). Originally, “accuracy” referred to the technical quality criterion of match-rate in classification, i.e., the relative number of correctly predicted classes. Its use, however, has overwritten that understanding with a more abstract interpretation as statistical comparison of machine learning system outputs with their desired counterparts.

Accuracy also is a central compliance measure within the EU AI act. For lawyers, accuracy is a quality certification that guarantees performance with respect to an intended purpose. The intended purpose is crucial: it makes no sense to regulate automobiles in terms of their use as medical interventions; their intended use is for personal transport. It explicitly takes into account the role of data generation, the importance of the transferability and robustness of accuracy to new application scenarios, and the harms and effects on society and individuals. Nevertheless, accuracy remains undefined and vague. The lawyer strives to guarantee accurate algorithms in the same way she strives to guarantee safe cars.

Given its centrality to both the legal and the technical community, one might expect that the machine learning community can “simply” explain to lawyers what accuracy is. It’s *not* that simple. Accuracy acquired new meanings, new facets and new expectations when it left the inner circle of machine learning. The community can not give a simple answer to the question: **what is accuracy?**

In this paper, we use “accuracy” as a case study for understanding interactions and frictions that arise when the technology of machine learning meets society. We claim that neither the technical nor the legal community have the ability to authorize a single universal notion of accuracy. Instead, we highlight how accuracy is negotiated at the boundary. The friction between the expectations set on “accuracy”, the use of “accuracy” and the aims of “accuracy” reveal a constructive potential, which leads to a set of recommendations to both the communities legal and technical. In other words, we use “accuracy” as a case study for identifying translation gaps, misunderstandings and needs at the intersection of the technological world of machine learning and the social world of law.

1.3 Contributions

Rather than considering what accuracy *is*, we focus upon what accuracy *does* in both technical and legal domains.

- We identify *E-accuracy* (Section 3 E) as a statistical, formal and empirical way of defining the relative performance of a machine learning method in comparison to others. The comparison is used to compete for high accuracy.
- We elaborate *L-accuracy* (Section 3 L) as a performance compliance measure, with the normative aim to prevent harms. Validity concerns about accuracy turn out to outweigh concerns about the methods of measurement. We provide detailed doctrinal analysis in Appendix L.
- The result of the parallel presentation is a structured comparison of the dimensions of friction (Section 4): (a) *Nature of accuracy*. Is accuracy formal and empirical or rather normatively and teleologically oriented towards the intended purpose? (b) *Notion of performance*. Does accuracy define performance or is part of it? (c) *Scope of validity*. To which extent is validity necessary and possible, for comparison or for use-cases with risks of harm? (d) *Ends*. Is accuracy used as an internal objective or rather an external certificate? (e) *Statisticalness*. Is the focus of accuracy based on populations or individuals?
- The named frictions offer ten, concrete, constructive recommendations, see **R1–R10**. These recommendations address the law community: We advocate for the reporting of accuracy with a focus on the justification of the use of metric and data, how the target of the AI system relates to the construct of intended purpose. The report, so we argue, should furthermore include baseline comparisons and interventional studies dependent on the context. These recommendations as well address the machine learning community: There are new (technical) ways of ensuring the validity of accuracy measurement demanded which go beyond comparative performance. Furthermore, the machine learning community offers tools to approximate “individual accuracy” which can be helpful from a legal perspective. Finally, some of the recommendations address both communities: We need to accept there is no end to the sense-making loop when bridging formal and normative definitions of accuracy. We should be aware of the limits to regulating statistical systems which view the world through aggregates of quantities, and the side-effects of this statistical stance which is promoted in machine learning, and reified in the EU AI act.

2 Accuracy as a Boundary Object

“Legal and political institutions lead, as much as they are led by, society’s investments in science and technology (Jasanoff 2004).” The technologies developed by AI-engineers are now used at scale in the wild. The intensified confrontation of the technology with society leads to a set of interactions in which society and societal norms are shaped and technical interests and requirements are redefined and co-produced (Jasanoff et al. 2004). This back and forth has

entangled and enriched concepts such as “accuracy”. To disentangle and “squeeze the juice” out of “accuracy”, we treat “accuracy” as a *boundary object*, shed light on it from two angles, and use it as a case-study to elaborate more generally on the interplay society and technology. In this section, we give background to our approach.

A *boundary object* is an “analytic concept of those scientific objects which both inhabit several intersecting social worlds [...] and satisfy the informational requirements of each of them” (Star and Griesemer 1989). Boundary objects are subject to “interpretative flexibility” and thus diverging sense-making across different groups (Star 2010) — in this case lawyers and engineers / scientists. Boundary objects function as an enabler for “different groups to work together without a consensus” (Star 2010). We keep dissensus about the meaning of accuracy throughout the entire study.²

The boundary object “accuracy” has not emerged all of a sudden in-between AI and law. It has rather undergone a transformative development, being stretched by the expectations set on it from a variety of different angles. A stylized history of “accuracy” in machine learning and legal scholarship follows: We arbitrarily declare the starting point of “accuracy” with the adoption of “classical accuracy”, i.e., rate of correct classifications, in machine learning, e.g., (Hughes 1968; Duda and Hart 1973).³ It initially appeared in the AI Ethics Guidelines of the EU Commission’s High-level Expert Group of AI as a quality of a trustworthy AI (HLEG 2019). However, it quickly morphed into the more sparsely worded legislative requirement in EU’s White Paper on regulating AI (European Commission 2020). While included already in the first draft of the EU AI Act (European Commission 2021), it lacks definition. It appears that the lawmakers referred to “accuracy” in the context of the AI Act, apparently presuming that the technological community knew what accuracy is. Subsequently, “accuracy” gained in relevance and developed to a more abstract term in the machine learning community, e.g., (Gupta and Lantaigne 2020). And, the circle continues. Our precis is stylized, but shows the entanglement, mutual interaction, and the difficulty of separating a legal versus technical understanding of “accuracy.” We will cut the boundary object “accuracy” roughly following the line of machine learning and legal scholarship. We will use *E-accuracy* to refer to the technical, machine learning conceptualization, and *L-accuracy* for the legal conceptualization. Whether the cut, which we propose, perfectly separates the understanding does not matter.

We illustrate how the stylized origin story of “accuracy” is resolved in the constant dialogical sense-making and conceptualization between a technical and legal community in Figure 2. External influences further shape “accuracy”. These effects will be touched only marginally in this work.

²This is a feature, not a bug.

³The use of the term in this context likely goes back to the precision-accuracy divide known to measurement theory (Tal 2020). Simultaneously, the closely related receiver operator curve (ROC) gained renewed interest in signal detection theory (Peterson, Birdsall, and Fox 1954). Even though “accuracy” was used within the context of classification, it was not simply defined as the mathematical definition Eq. (1) we provide in the Appendix.

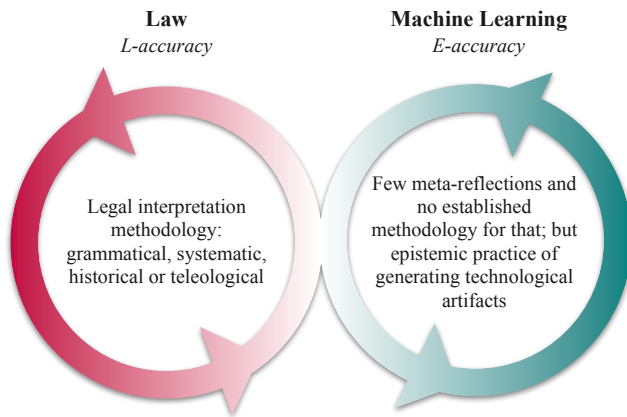


Figure 2: The stylized process of sense-making in the context of the term “accuracy”. The conceptualization of the term “accuracy” in law (L-accuracy) and the conceptualization of the term “accuracy” in machine learning (E-accuracy) follow their own logics or dynamics. We are particularly interesting in the mutual interaction between the sense-making loops, which is caused by frictions in the understanding of the term “accuracy” (marked as white transition between the circles).

The tensions arise because of the distinctive epistemic practices in both worlds. This becomes evident upon examining the question “what is accuracy?” Machine learning has no established methodology to answer that. Meta-reflections are largely absent within technical research, because they do not produce new technological artifacts. However, there exist commonly agreed upon epistemic practices for the development of technology, in which “accuracy” plays a vital role (Section E.3). Hence, within this realm we elaborate on how accuracy is used, and approach a meta-reflection on accuracy by circulating around the technical literature. In particular, we mainly take into consideration existing accounts with a socio-technical focus (Section 3 E).

Accuracy represents a technical concept introduced into law (Koivisto, Koulu, and Larsson 2024). Such concepts, including “algorithmic bias and fairness” and “explainability” carry normative meanings also in the technical community, which may or may not find their correspondence in law (Kaminski 2021; Weerts et al. 2023). For example, explainability techniques are viewed insufficient to meet the normative standards that the law sets, or should set for algorithmic decision-making (Edwards and Veale 2017; Cofone 2026).

Technical terms represent linguistic transplants to law. When left undefined, they create uncertainty about their interpretation (Koivisto, Koulu, and Larsson 2024). One the one hand, the technical community may give multiple meanings to the given concept (Weerts et al. 2023; Koivisto, Koulu, and Larsson 2024). On the other hand, when integrated into law, the concepts become subject to the practices of legal interpretation (Koivisto, Koulu, and Larsson 2024). The requirement for “accuracy” in the AI Act, must be seen through the regulatory objectives set for the AI Act. Sim-

ilarly to compliance criteria present in other product safety regulations, the conditions for high-risk AI are reflective of a metaregulatory approach. The EU legislator enacts relatively abstract criteria for high-risk AI, which are to be specified by harmonized standards set by standard-setting organizations CEN and CENELEC (European Commission 2025). In the absence of, or instead of, the more detailed standards, the providers may choose their own way to comply with the requirements (Pouget, Holtman, and Emborg 2025).

This means providers – i.e., computer scientists – interpret and give meaning to the notion accuracy in the routine compliance with the AI Act. However, when incorporated into legal act, the term accuracy becomes part of law (Koivisto, Koulu, and Larsson 2024). Lawyers and the judiciary will define the meaning of the term based on established methods of legal interpretation. Common methods of interpretation are (a) grammatical (focus on the wording), (b) systematic (focus on the system of rules, i.e., the AI Act and other EU law), (c) historical (preparatory works; earlier related pieces of law), (d) teleological (focus on the objectives of the AI Act and its impacts) (Möllers 2022). Critically, this legal maneuver might uproot terms from one habitat, and appropriate them to law – with unpredictable consequences for future legal interpretations (Koivisto, Koulu, and Larsson 2024).

Through E-accuracy and L-accuracy we will observe two epistemic disciplines with self-reproduction of their knowledge (Luhmann 1992). Both disciplines make sense of “accuracy” in their own terms. For instance, machine learning accepts “accuracy” only to be expressible in formal, technical language, otherwise, it is not of value for the development of technological artifacts (Section E.2). On the other hand, legal scholars seek for consistency of the use of “accuracy” across legal text. That results in an understanding of “accuracy” as a normative, teleological concept (Section L.4). The surprising observation is that both sides can easily refer to their sense-making as the “right” and “true” one, a phenomenon called “self-vindication” by Ian Hacking (Hacking 1992): It is truth what we found, because we used a valid method; the method is valid because it leads to the truth. In this text, we accept both the “truths”, the “true” meaning of accuracy as E-accuracy and the “true” meaning of accuracy as L-accuracy.

2.1 Organization of the Paper

The remainder of the text is structured closely to the methodological approach laid out above. Below we present E-accuracy and L-accuracy in parallel threads in a double-column format (Section 3 L and Section 3 E). In the main paper we summarize the claims. In the supplementary material (Appendix L and Appendix E of the full version (Derr, Wernick, and Williamson 2026)), we give the “proofs” for those claims. Both threads are independently readable, but mutually complementary. The tensions created in these sections are then resolved, evolved and convolved in Section 4. The two “truths about accuracy” reveal requirements for action, e.g., in standardization and technical research, and fundamental limits in the reconciliation process. We conclude by taking a satellite-perspective on the relationship between machine learning and law.

3 L What is L-Accuracy? AI Lawyer's View on Accuracy

Being the first comprehensive effort to regulate AI, the EU AI act aims to improve the functioning of the internal EU market and promote the uptake of human-centric and trustworthy AI (AIA rec. (1)). A key compliance requirement within the Act is “accuracy”, a concept which is alluded to also in the context of documentation. Despite its prominent role, “accuracy” is not defined. We outline below how and in which context “accuracy” is used within the EU AI act.

Introduction to **regulation** of (computational and statistical) **products** (L.1)

2024 the EU AI act entered into force. The EU AI act is a product safety regulation concerned with AI systems. It is based on risk classifications which depend on the use of the AI system. In particular, high-risk AI systems, e.g., when used in healthcare or education, need to fulfill a series of compliance requirements.

L-accuracy is a **legal compliance** measure (L.2)

Among other compliance measures, high-risk AI systems need to “achieve an appropriate level of accuracy” following Article 15 of EU AI Act. More specifications and demand on the reporting of accuracy are scattered in the Act. In fact, “accuracy” is not explicitly defined. However, it can be considered as part of “performance”, the ability of an AI system to meet its intended purpose.

L-accuracy **certifies** the **fitness-to-purpose** of machine learning **systems** (L.3)

The EU AI act aims to certify the development process of AI systems. Part of this certification is accuracy. In particular, accuracy is to be understood with respect to the intended purpose of the AI system. The intended purpose, within the scope of high-risk use cases, can range from predicting risk of school drop-out to interventional scenarios such as allocating support based on drop-out risk.

L-accuracy is inherently **normative** and used for **risk mitigation** (L.4)

With the requirements around accuracy, the EU AI act aims to mitigate risks for health, safety and fundamental rights. Risk mitigation, so it is argued, is achieved by ensuring accuracy “throughout the lifecycle”. That reading makes L-accuracy *inherently* normative. In the anticipating attempt to reconcile the normative with a technical understanding, the AI Act highlights, beyond the transparency of the accuracy measurement, the centrality of validity (arguments) of the accuracy measurement.

L-accuracy **exists** to **protect** subgroups and individuals (L.5)

Generally, law has a focus on individuals and individual's rights. More specifically, accuracy is explicitly mentioned in the context of subgroups and individuals within the EU AI act. Accuracy can serve to protect them.

3 E What is E-Accuracy? ML Engineer's View on Accuracy

Accuracy is so central and standard to machine learning that the question what it *is* feels inappropriate. Nevertheless, carving out the facets of E-accuracy will serve several purposes here. We make explicit what is unarticulated, implicit knowledge in machine learning. In doing so, we emphasize aspects which might seem marginal to the machine learning community, which, however, turn out to be fundamental anchors for tensions with respect to L-accuracy.

Introduction to **engineering** of (computational and statistical) **technologies** (E.1)

Much of machine learning has become engineering, that is the construction and maintenance of a technology. In particular, machine learning creates computational systems for generating predictions, recommendations, or other artifacts. Critically, those systems are statistical, and hence are based on aggregates of data.

E-accuracy is a **formal and empirical** measure (E.2)

Classically, accuracy is understood as the match-rate between predicted and actual classes. More generally, accuracy involves two main ingredients: metric and data. Accuracy is the formal comparison (based on the metric) of desired and generated outputs, e.g., classes, rankings, regression values, on an aggregate of instances (based on the data). Hence, accuracy is outcome-focused, formal, empirical and statistical.

E-accuracy **defines** the **performance** of machine learning **models** (E.3)

Accuracy is used as an abstract performance objective and even defines performance. The closer to deployment, the more the limits of accuracy are acknowledged even though accuracy still gives decision support on which model to use. Missing validity, e.g., because of mismatches in data, metric, constructs and interventions, limits the scope of accuracy in defining model performance.

E-accuracy is inherently **statistical** and used for **ranking models** (E.4)

Accuracy is *inherently* statistical (beyond E.2), as it favors statistical methods over other approaches, because those methods are optimized for exactly that goal (E.3). Currently, this optimization progress is measured in *benchmarks*, a type of standardized accuracy measurement, which is based on ranking models. The use of accuracy in machine learning (benchmarking) only requires such relative comparisons. The validity of such rankings beyond the measurement scenarios are backed, but hard to hold beyond comparisons.

E-accuracy is **poorly suited** to subgroups and individuals (E.5)

Individual accuracy, i.e., the claim that a statistical system is accurate on a single, unobserved instance, is a hardly justified and even oxymoronic concept. Statistical systems, more generally, act on the level of aggregates.

4 From Clashes to Recommendations

In order to understand and mediate the resulting tensions of AI in socio-technical reality, we picked out “accuracy” as a recurring topos in legal text on AI, as well as scientific and engineering literature. We now elaborate on clashes of expectations and conceptualizations at the boundary. Despite the destructive connotation of “clash”, we are convinced that the themes of friction and disagreement are productive starting points for both communities. In particular, we reveal the necessity for technical progress, recommendations for the ongoing standard-setting process for accuracy (CENELEC 2026) and fundamental limits to technical developments and remaining challenges which ask for changes in law, marked as recommendations. The structure of this section is given in Table 1. To facilitate understanding we introduce a running example.

***Example.** Imagine an automated AI-tool “E-Grad” for grading English exams in high-schools. This tool likely classifies as high-risk AI system under the EU AI act. Hence, it needs to fulfil the specified demands on accuracy. Furthermore, there is little additional and intertwining regulation relevant to educational scenarios different to healthcare.*

Our recommendations, especially (R4), (R6) and (R9), target the ongoing standardization process for high-risk AI. To our knowledge, the issues highlighted have not been visibly addressed in the standardization process, which is at an early stage and opaque. The standards are prepared by CEN/CENELEC Technical Committee JTC 21. The sub-project for drafting of the accuracy standard (prEN 18229-4) began in June 2026 within its Working Group 4 on Foundational and ethical aspects of AI (GENORMA 2026; CENELEC 2026; CEN/CENELEC 2026). In addition, compliance criteria for accuracy are likely to be informed by standards for computer vision (prEN 18281) and natural language processing (prEN ISO/IEC CD 23282), the revised ISO/IEC DIS 4213 standard for AI performance evaluation and the standards for other high-risk AI obligations, such as transparency (Digital 2026), see also (Marin et al. 2026).

4.1 Nature of Accuracy

While E-accuracy is a formal and empirical concept, which can be written down mathematically and computed on data (E.2), L-accuracy remains elusive with its aim of use within the EU AI act as a centrally defining feature (L.4). As the EU AI act acts on actual technology, a translation of the formal, empirical concept E-accuracy into its teleological counterpart is necessary. Marin et al. (2026) elaborate extensively on the choice of accuracy metric, the appropriate threshold and the test data. That entire process can be understood as the attempt to equip formal definitions of accuracy a teleological meaning. That particularly requires the effort for proper justifications of the choices made (Section 4.3).

The reconciliation of the natures of accuracy is (successfully) underway, e.g., standardization (European Commission 2025), new forms of evaluations like arenas⁴, and research developments (Marin et al. 2026). Without question-

ing the progress, we want to emphasize a lesson of the confrontation of E-accuracy and L-accuracy. The different natures of accuracy, which are tightly linked to the languages of the communities, are in principle *not* translatable. L-accuracy defeats a full formalization and expression as an empirical measurement, because by its teleological definition it needs to follow the vague and mobile aims, such as as mitigating risks (Section L.4). Conversely, E-accuracy represents a static, momentary status. That discordant note, however, need not concern us.

(R1) *We have to accept that the search for formal and empirical representations of accuracy within the context of legal compliance will continue indefinitely.*

For the better, that way accuracy respects the dynamics of our society and values and the friction motivates the development of new technical measures.

4.2 Notion of Performance

From the perspective of L-accuracy, “performance” is an overarching term, to which accuracy, but as well robustness or cybersecurity, belongs to (L.2). In particular, performance, including accuracy, needs to be considered with respect to the intended purpose of the AI system (L.3). That holistic reading of performance and its pragmatic reading of accuracy, i.e., accuracy with respect to the intended purpose, is reflected in the limitations of E-accuracy (E.3). While we argued that E-accuracy defines performance, it does so only within a limited scope. Critically, E-accurate machine learning models do not translate into successful applications for the same reasons for which E-accuracy does not translate into L-accuracy. The mismatches given in (E.3) can be understood as translation difficulties from E-accuracy to L-accuracy.

Accuracy metric mismatch is the problem that the choice of accuracy metric needs to capture performance semantics. Hence,

(R2) *a report of an accuracy metric needs to argue why it is semantically meaningful.*

Data set mismatch is the problem that the choice of data on which accuracy is measured needs to be “representative” and meaningful for the deployment context. That is,

(R3) *a report of a data set needs to argue why the accuracy measurement on this data set is valid in a relevant scope, i.e., “throughout the lifecycle”, and respects robustness in deployment changes.*

Target-construct mismatch is the problem that the target which the AI systems aims to predict, recommend, generate is only in weak relationship to the abstract construct of intended purpose. It follows that

(R4) *a report of accuracy needs to specify the AI system’s target and its relationship to the intended purpose.*

The named mismatches are in fact concerns of missing validity (cf. Section 4.3), see (Marin et al. 2026) for an elaboration on the first two challenges.

***Example.** Accuracy of E-Grad measured as agreement of grade by tool and teacher(s) sounds reasonable. R2: But*

⁴<https://arena.ai/leaderboard>

Sec.	Aspect	L-Accuracy	E-Accuracy	Recommendations
4.1	Nature of accuracy	Teleological (Towards intended purpose)	Formal and empirical	R1
4.2	Notion of Performance	Part of performance for intended purpose	Defines comparative performance	R2, R3, R4, R5
4.3	Scope of Validity	For mitigation of risks in actual deployments	For comparison	R6, R7
4.4	Ends	External certificate (to get better social outcomes)	Internal objective (to improve and choose method)	R8
4.5	Statisticalness	Focus on individual stance	Population based statistical stance	R9, R10

Table 1: Aspects along which E-accuracy and L-accuracy clash.

why does agreement matter? It reflects teacher’s gradings, but shouldn’t it rather matter that all potential spelling errors are found and considered for the grade? R3: Who is the teacher with whom the tool is compared with? From the same school? Averaged over multiple teachers from multiple schools? R4: Is the intended purpose of the AI system to reduce the workload of the teacher while the English skills of the students remain well-captured and why does the grading matter then?

A Plea For Decision-Accuracy The EU AI act implicitly assumes that accuracy is a performance measure which applies to decisions produced by AI systems (L.2). In addition, the EU AI act does not make any restriction on whether the intended purpose of an AI system is interventional or not, i.e., whether the success of the AI system deployment needs to take into account changes happening due to the deployment (L.3).

Example. *If E-Grad’s grades are used to decide which student to support by an exchange program, the agreement of grading between tool and teacher are less relevant than the success of assigning students to the exchange program. In this case, what matters substantially are the effects on the English skills and whether the “correct” group of people was sent to the program.*

We suggest to make an interpretative move towards *decision-accuracy* when there is an interventional purpose of an AI system. That is, we ask for accuracy to be understood as measuring the accurate assignment of treatment with maximal causal effect. That requires a lot of the machinery of quantitative social science: (a) specification of the desired causal effect, (b) randomized data for evaluation, and (c) a model of effect (Angrist and Pischke 2009). The standard understanding of accuracy in machine learning is far from any such interpretation and practice. Nevertheless, under certain circumstances the standard understanding of accuracy is a good proxy for decision-accuracy (Pearl 2021; Imbens and Athey 2021; Fernández-Loría and Provost 2022).⁵ However, if those conditions are not met, E-accuracy becomes a meaningless measure for the performance with

⁵Do we need counterfactual predictions of outcomes, e.g., potential outcomes (Rubin 1974) or counterfactual loss (Koch and Imai 2025), and randomized (interventional) data to circumvent

respect to an interventional intended purpose (Barabas et al. 2018). The harmonization legislation of certain high-risk AI systems (Annex I(a)) already require interventional studies: MDR requires the proof of causally verifiable clinical benefit (Art. 2 (44)(52)). In comparison to (European Parliament and Council 2017), EU AI act’s notion of “performance” makes not reference to causality (Art. 3(18)). However, the overarching objective of risk-mitigation and risk profiles of Annex III (AI applications) support interpreting the requirement of accuracy to include “decision accuracy” and interventional studies. It should be acknowledged that interventional studies may require real-world testing within (Art. 57) or outside AI regulatory sandboxes (Art. 60).

(R5) Standardization for the EU AI act needs to address accurate treatment assignment with causal effects.

Tools such as randomized control trials, differences-in-differences, etc. (Huntington-Klein 2021; Angrist and Pischke 2009) should be part of the default canon of accuracy measurement techniques. From a technical side, we’d profit from a clearer understanding of the proxy relationship between E-accuracy and decision-accuracy, e.g., (Fischer-Abaigar, Kern, and Perdomo 2025; Fischer-Abaigar et al. 2026) and how accurate treatment assignments can be efficiently established (Casacuberta and Hardt 2026).

4.3 Scope of Validity

For E-accuracy, the scope of validity is set by its use as a competition marker (E.4). Valid relative comparison of methods needs to be possible beyond the accuracy measurement setup. In contrast, L-accuracy is rather more focused on its scope of validity “throughout the[ir] lifecycle” (AIA Art. 15(1)) and under “known and foreseeable circumstances” (AIA Art. 13(3)(b)(ii)) for mitigating risks in actual

the problems mentioned? Partially. First, it is clear that causal effect estimation, or counterfactual predictions of outcomes, is sufficient for “accurately assigning treatments to maximize causal effect” (Fernández-Loría and Provost 2022). However, it is not necessary. There are conditions under which accurate outcome predictions (prediction-accuracy) can be useful for interventions: (a) when outcomes and effects are correlated, (b) when the gain in estimation accuracy is strong in comparison to identifying the best treatment (c) when predictions need rank consistency but not absolute accuracy (Fernández-Loría and Provost 2022).

deployment (L.4). There is a mismatch in the scope of validity, which let's us join the chorus of criticism regarding current AI evaluations, e.g., (Hutchinson et al. 2022; Eriksson et al. 2025).

Judgements about “goodness” of AI systems require contextualization. It is certainly true that with benchmarks we observe intradisciplinary progress and that “machine learning runs on a weaker currency: comparisons and rankings” (Hardt 2025). But, concretely, for compliance to the EU AI act, or more abstractly, for following normative social requirements, machine learning scholarship needs to serve the interests outside the discipline for extended scopes of validity of accuracy measurements with a focus on harmlessness and social benefit. More concretely, the mismatch in the scope of validity brings the following two aspects to light.

A Request for Baseline Reportings in Legal Compliance to Extend Validity and Demonstrate Alternatives Machine learning scholarship observed that comparative accuracy measurements are substantially better justified to be valid beyond the measurement scenario than direct scores (Section E.4). In other words, model rankings defined by their accuracy often transfer reliably to new test scenarios, while the actual accuracy scores do not. Thus we advocate **(R6) report baselines and comparison systems as part of the standards for EU AI act compliance.**

Baselines can extend the validity of accuracy measurements.

Furthermore, baselines demonstrate alternatives to the use of an AI system. In (Wang et al. 2024) the authors advocate that an AI system needs to prove its advantages along a set of dimensions over existing methods to justify its use. For instance, in (Dressel and Farid 2018) the authors show that non-experts and extremely simple statistical methods perform as well as a presumably complicated recidivism score prediction tool COMPAS. The strand of literature on the “simpler” alternative, e.g., following (Letham et al. 2015; Semanova, Rudin, and Parr 2022), substantiates this approach. Hence, reporting the baseline is crucial as it enables deployers and users to take a substantiated choice of whether and how to use an AI system.

Example. *Accuracy reporting for E-Grad via a benchmark which consists of annotated, graded English exams at UK high-schools should be accompanied by baselines. That is, the report should contain benchmark results of, e.g., comparable AI systems or selected teacher re-gradings to substantiate the accuracy claims.*

The introduction of comparison systems comes with the burden to choose a comparison system. Which comparison system should we consider? Rule-based systems (Johnson and Zhang 2022)? Simple statistical baselines (Letham et al. 2015)? Humans? Competitor systems? The provider needs to take the burden of defining and justifying a reasonable comparison system, which at least is (a) relevant for the intended purpose, (b) has shown accuracy achievements and (c) is distinct to the process of data set creation. Those conditions seem trivial, but can, depending on context, exclude certain choices. For instance, humans are often inaccurate (Dawes, Faust, and Meehl 1989) and not optimized for cer-

tain abstract accuracy metrics (Recht 2025), plus they are often central players in the data creation pipeline. For instance, if the labels for images were given by certain human annotators, one needs a specific argument why a comparison of a machine learning system should be pursued against a new set of human labelers. Hence, it is easy to exclude the possibility that there is a naive, universal choice of comparison system. Rather, the choice of a comparison system, as its justification, is bound to the context and purpose of the AI system.

AI Systems Lack a Diverse Set of Certifications On a more abstract level, the challenge of validity of performance conditions such as “accuracy” is not specific to AI systems. For instance, proving a drug’s effectiveness by clinical trials is often based on statistical tests which are claimed to be valid beyond the participating cohort. That inductive step from trial cohort to unknown patient is held to be justified, e.g., because the used randomized control trial safe-guards against hidden confounders. Beyond the statistical test, the drug’s effectiveness is certified, e.g., by bio-chemical tests and knowledge of effect mechanisms and the structure of the drug. Those certifications are even legally mandated, e.g., (European Parliament and Council of the European Union 2001, Article 8(3)(i)). With the diverse arsenal of methods, the scope of validity can be refined, e.g., if the drug’s mechanism is known to rely on certain enzymatic complexes which diabetics don’t share, then the drug is known to be useless for such patients. Hence, in which way is the validity challenge for accuracy of AI systems special?

The validity challenge is particularly difficult for AI systems, because (a) AI systems are often treated as a general-purpose technology⁶ and (b) AI systems of current interest are statistical.

Regarding (a) the variety of use-cases demand a diverse set of certifications of validity which needs to be represented in standards and regulation.⁷ Regarding (b), the statisticalness is not only a property of the AI system, but as well a central component of the problem formulation.⁸

⁶The EU AI act is neutral to the set of (intended) purposes (AIA Art. 2, Art. 3(1)) and even defines “general-purpose AI models” and “general-purpose AI systems” (AIA Art. 2(63)(66)). The theoretically-inclined machine learning scholars make no semantic commitments regarding the inputs and outputs of their treated systems. The business-oriented community (over)claims no limits to the use-cases of machine learning products (Widder and Hicks 2024).

⁷There is always some margin for refinement left here for any product safety regulation. For instance, what is the exact use of a medical device or a construction product and the relevant ways how they perform? However, the scope is still much narrower as such products are already defined through their abstract purpose, e.g., (European Parliament and Council 2017, Art. 2(1)) and (European Parliament and Council 2011, Art. 3(1)).

⁸In fact, the EU Council wrote in the version of the EU AI Act from November 25 2022: “The range of problems addressed by machine learning typically involves tasks for which other approaches fail, either because there is no suitable formalisation of the problem, or because the resolution of the problem is intractable with non-learning approaches” (Council of the European Union 2022).

It follows that the compliance of such a statistical system needs to heavily rely on statistical measures, such as accuracy. There is no causal, mechanistic, social, biological, chemical, physical, etc. way to argue that a certain AI system is accurate. Drug effectiveness, in contrast, can be corroborated, e.g., by bio-chemical knowledge; there is another means for refining the scope of validity.

(R7) Machine learning systems, hence AI systems, lack a diverse set of such certifications, beware of that.

Example. E-Grad as AI-tool based on machine learning methods is a statistical technology. Patterns of text (English exam) to number (grade) are distilled from training data sets. Roughly, the patterns are distilled based on empirically testing a set of possible mappings and choosing the best among them. The choice of tested mappings is rather broad and general, often large neural networks. There exist little to no arguments beyond “the chosen mapping is the statistical best” which can be used to corroborate that E-Grad as a tool will work as expected under varying conditions.

4.4 Ends

E-accuracy is used as an *internal* objective within the machine learning community to improve and choose methods (E.3 & E.4). L-accuracy certifies *externally* the quality to users, deployers and other affected persons (L.3 & L.4). Despite the distinct uses of accuracy, “being accurate” is positively connoted. Everybody wants accurate AI systems, but in which way?

The “iron rule” of machine learning following Hardt (2025) is the competition on benchmarks, an institutionalized measurement of accuracy. In its modern use, there are almost no rules on how to compete, except maybe for training or memorizing new methods on test sets (Xu et al. 2024). Hence, “gaming” the competition is what happens (cf. (Blum and Hardt 2015)), but is part of the definition of progress (cf. (Orr and Kang 2024)) and is hence construed as a desirable feature.

Very differently, accuracy as compliance measure is a burden to the provider. There is a necessity for being accurate which is (a) not fully intrinsically motivated, but rather implied by the regulator, and (b) might not be aligned with respect to one’s own intentions. There is likely an interest by the legal stakeholders to be accurate, but to demonstrate harmlessness, certain accuracy measurement might be needed which are against the interest. A simple example is the classification accuracy of a diagnosis tool: provider, deployer and patients might agree that a low false negative rate is desirable, i.e., not many actually sick patients are diagnosed as healthy. However, a low false positive rate might not be in the interest of the deployer, hence of no interest to the provider, who will sell more treatments, but in the interest of patients who do not want to undergo unnecessary treatment. So, providers can safeguard themselves against the accuracy compliance measure by carefully designing an irrelevant but easily fulfilled accuracy metric.

(R8) We need to anticipate that “accuracy-hacking” will matter and develop remedies against it

(cf. fairness-hacking (Black, Gillis, and Hall 2024; Simson, Pfisterer, and Kern 2024; Meding and Hagendorff 2024)).

4.5 Statisticalness

In contrast to E-accuracy and machine learning methods, which are statistical and thus focused on the aggregate, L-accuracy refers to individuals (L.5). This tension is particularly problematic, as we argued that from the perspective of E-accuracy “individual accuracy” is oxymoronic (E.5).

Example. If E-Grad has 95% agreement with a teacher’s grades on a set of English exams, this accuracy report does not and cannot tell anything about whether one individual exam (even in test set) was accurately graded. What can be said about the accuracy of E-Grad on an unobserved English exam?

Practically Reconciling the Statistical and the Individual Stance A legitimate step forward is to interpolate between individuals and the overall aggregate. The EU AI act explicitly mentions not only individuals but as well “subgroups” (Section L.5). In that regard “individuals” are the limit of a spectrum of finer and finer subgroups. Hence, while “why is the AI system accurate on this individual” is hardly justified (E.5), the question “why is the AI system accurate on this subgroup” at least partially relieves the burden of the validity argument.

In this regard, it might be helpful to demand for instance-wise reports of AI system’s outputs and their desired counterparts, e.g., along the line of (Burnell et al. 2023). That way, subgrouping along different post-hoc specified categories is possible. Together with tools from fair machine learning, which often are concerned with worst-case performance on subgroups (Zafar et al. 2017; Pleiss et al. 2017; Hébert-Johnson et al. 2018; Donini et al. 2018; Williamson and Menon 2019; Loi and Heitz 2022), and the analysis of Rashomon sets, i.e., sets of machine learning models with equal performance (Frohnepfel et al. 2026),⁹ valid approximations of “individual accuracy” will be practically possible. Hence, by refraining from the ideal of individual accuracy,

(R9) we should use implementable methods to obtain justified estimates for subgroup accuracy or indication of failure of individual accuracy.

EU AI act Codifies Statisticalness of Our (Social) World

Besides the challenges around individualization, the EU AI act is an example of codifying a statistical stance into law, which itself usually is closer to an individual stance. Due to the demand of accuracy, which has attached a statistical interpretation, statistical methods such as machine learning techniques gain a new type of foothold on a social level. The development-internal justification of machine learning systems, i.e., good statistical accuracy, attain a new relevance beyond the developing community. The EU AI act

⁹Concretely, the authors of (Frohnepfel et al. 2026) suggest to study the disagreement of machine learning models. Disagreement can be an indicator for the invalidity of of an individual accuracy measurement beyond the measurement context.

hence reproduces *patterns of “performance”* shaped by engineers and developers. Yes, the EU AI act *cannot* be ignorant towards the technology it wants to regulate. Nevertheless, it performs and reifies “accuracy” as a central goal for the technology of machine learning (Hildebrandt 2020, §2.1.2)(MacCormick and Weinberger 1986). In that regard, one might consider it a literal *law* of large numbers.

The problem is that the patterns of “performance” might not be congruent with a broader societal consensus. From a machine learning perspective, measuring everything in accuracy aligns well to its own principles. From a broader perspective, the relevance of accuracy (Section E.3) in particular, and the representation of the (social) world in statistical terms in general, can be questioned. For instance, Eaglin (2019) argues that statistical systems for risk assessment in criminal justice restrict our ways of talking about justice. It is hard to imagine what the exact consequences of the choice of “accuracy” as central statistical compliance measure within the EU AI act are. But,

(R10) we should consider the side-effects of a statistical stance to the world, promoted by ML and legislation.

Example. Imagine the EU AI act legitimates the use of the AI-tool E-Grad. Hence, it promotes that statistical grading systems are accepted in our society. Do we want grading to be procedure governed by quantified data comparison? And how does that relate to the educational purpose of grading in the first place?

5 Lessons Learned – A Satellite Perspective

We examined two parallel threads of sense-making to “what is accuracy” from a machine learning and a law perspective in the above. Actually, the first meta-lesson we learned is the productive step from “accuracy” as a noun to “accuracy” as a verb. We more productively asked “what accuracy *does*.”

Both perspectives project their own needs, expectations and limits on accuracy, but agree that there is some concept of interest. This pattern is not specific to “accuracy” at all. Terms and concepts such as “robustness” (Freiesleben and Grote 2023; Nolte, Rateike, and Finck 2025), “explanation” (Bordt et al. 2022; Bordt, Raidl, and Von Luxburg 2025), “interpretability” (Molnar et al. 2020), “privacy” (Bailie 2025), “fairness” (Barocas, Hardt, and Narayanan 2023) are created and negotiated at the boundaries of machine learning. The self-referencing circles “what is ...” within the technosphere break apart the more the technology leaves the laboratory. Our study of “accuracy” holds some general lessons for all those terms and concepts.

The clash of non-overlapping interpretations of technical terms is inevitable. We even should not expect those clashes to settle (cf. Section 4.1). Machine learning is restricted by its need for formalization and mathematization. However, that creates a force which easily clashes with other (social) perspectives. For instance, law is restricted by its attempt to consider competing social interests which often results in ambiguity and vagueness. Ambiguity and vagueness are the enemies of proper formalization (cf. debates on fairness metrics, e.g., (Barocas, Hardt, and Narayanan 2023)).

A second conflict arises along the line of statisticalness. Machine learning is based on statistics, while legal or other perspectives often are concerned with the individual. The statistical stance versus the individual stance will productively generate disagreement about interpretation (cf. see the debate on explanations, e.g., (Bordt et al. 2022; Bordt, Raidl, and Von Luxburg 2025)). Relatedly, while there is plenty of precedent in regulating new technologies, there is not much precedent in the regulation of *statistical* technologies. It now seems that it is not just philosophers but also lawyers who have to grapple with the question of how does the aggregate relate to the individual.

Finally, the internal logics of machine learning, the drive towards new methods (Sambasivan et al. 2021; Birhane et al. 2022), the solutionism (Reyes-Cruz et al. 2025), oppose the internal logics of other social spheres such as law, with an ideal of consistency (Hildebrandt 2020, §2.2.2). Unsurprisingly, non-engineering ways of thinking don’t need technical solutions for self-vindication. Still, the ways of thinking interact, and making sense of technical terms inevitably ends in a clash.

Even though clashes (will) occur, they (will be) are productive. The disagreement on meanings projects inside machine learning as well as outside. Exemplarily, the use of “fairness” got refined in machine learning (Barocas, Hardt, and Narayanan 2023) and sparked a debate beyond the technical community. **The clash of non-overlapping interpretations is productive and refines our understanding of the term.** Furthermore, such clashes repeatedly teach us the lesson to not put too much trust in abstract nouns with a naive belief in univocity.¹⁰

What remains open as an interesting and important avenue for future work, is the study of power structures in defining relevant terms and interpretations (Jasanoff et al. 2004; Koivisto, Koulu, and Larsson 2024). How are frictions utilized, by whom for which purpose? Near term, it is relevant to study whether terminologies adopting existing standards influence the notion of accuracy in the EU AI act’s standard setting process (ISO n.a.; Deng 2025), and how the JTC 21 working groups’ modular approach standard drafting for AI act Arts. 9-15 aligns with the holistic nature of L-Accuracy cf.(Marin et al. 2026; Vaciago 2026; Derr, Wernick, and Williamson 2026).

Acknowledgments

Thanks to Mara Seyfert for discussions; James Bailie, Riikka Koulu and Suvi Sankari for literature suggestions; and Emeline Banzuzi for feedback. This work was supported by the Carl Zeiss Foundation, the Kone Foundation (project “Long-term human rights risks of smart city technologies”) (AW); the Deutsche Forschungsgemeinschaft (EXC number 2064/1; Project number 390727645) (AW, RW) and the Tübingen AI Center (RD, RW).

¹⁰The difficulties that abstract nouns create is nicely demonstrated by Williams (2014) who grapples with examples such as “equality,” “individual,” “masses,” “society,” and “technology.”

References

- Acemoglu, D.; Autor, D.; Hazell, J.; and Restrepo, P. 2022. Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1): S293–S340.
- Aghion, P.; Jones, B. F.; and Jones, C. I. 2017. Artificial intelligence and economic growth. Technical report, National Bureau of Economic Research.
- Almada, M.; and Petit, N. 2025. The EU AI Act: Between the rock of product safety and the hard place of fundamental rights. *Common Market Law Review*, 62(1).
- Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Bailie, J. 2025. *Topics in Privacy, Data Privacy and Differential Privacy*. Ph.D. thesis, Harvard University.
- Barabas, C.; Virza, M.; Dinakar, K.; Ito, J.; and Zittrain, J. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*, 62–76. PMLR.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.
- Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review*, 104: 671.
- Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.
- Black, E.; Gillis, T.; and Hall, Z. Y. 2024. D-hacking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 602–615.
- Blum, A.; and Hardt, M. 2015. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 1006–1014. PMLR.
- Bordt, S.; Finck, M.; Raidl, E.; and Von Luxburg, U. 2022. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 891–905.
- Bordt, S.; Raidl, E.; and Von Luxburg, U. 2025. Position: Rethinking Explainable Machine Learning as Applied Statistics. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Burnell, R.; Schellaert, W.; Burden, J.; Ullman, T. D.; Martinez-Plumed, F.; Tenenbaum, J. B.; Rutar, D.; Cheke, L. G.; Sohl-Dickstein, J.; Mitchell, M.; Kiela, D.; Shanan, M.; Voorhees, E. M.; Cohn, A. G.; Leibo, J. Z.; and Hernandez-Orallo, J. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641): 136–138.
- Casacuberta, S.; and Hardt, M. 2026. Good Allocations from Bad Estimates. arXiv:2601.05597.
- CEN/CENELEC. 2026. Working Groups & Projects — CEN-CENELEC JTC 21.
- CENELEC. 2026. AI trustworthiness framework – Part 4: Accuracy, prEN 18229-4. Technical report, CEN.
- Cofone, I. 2026. Institutional Accountability and Legitimate Inference in Algorithmic Adjudication: Beyond Trustworthy AI. Available at SSRN 6516459.
- Council of the European Union. 2022. 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. General Approach 14954/22, Council of the European Union, Brussels. Interinstitutional File: 2021/0106(COD).
- Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science*, 243(4899): 1668–1674.
- Deng, A. L. S. 2025. Task-level AI Assurance. *Substack: AI regulation, standards and reality*.
- Derr, R.; Wernick, A.; and Williamson, R. C. 2026. Law of Large Numbers: Accuracy as a Statistical Measure for Compliance and Competition. arXiv. Forthcoming.
- Digital, T. 2026. EU AI Act Harmonised Standards Mapping. Summary of intended coverage by CEN-CENELEC standards.
- Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2796–2806.
- Donoho, D. 2024. Data Science at the Singularity. *Harvard Data Science Review*, 6(1).
- Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580.
- Duda, R. O.; and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. Wiley.
- Eaglin, J. M. 2019. Technologically Distorted Conceptions of Punishment. *Washington University Law Review*, 97: 483–544.
- Edwards, L.; and Veale, M. 2017. Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16: 18.
- Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 850–864.
- EU. 2026. Regulation 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) (Text with EEA relevance). Technical report, EU.
- European Commission. 2020. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust. White Paper COM(2020) 65 final, European Commission.
- 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. Proposed Regulation COM(2021) 206 final, European Commission.
- European Commission. 2025. COMMISSION IMPLEMENTING DECISION of 23.6.2025 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation as regards high-risk AI- systems in support of Regulation (EU) 2024/1689 of the European Parliament and of the Council and repealing Implementing Decision C(2023)3215. Commission Implementing Decision C(2025) 3871 final, European Commission, Brussels.
- European Parliament; and Council of the European Union. 2001. Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use. *Official Journal of the European Union*, 311: 67–128.
- European Parliament and Council. 2011. Regulation (EU) No 305/2011 of the European Parliament and of the Council of 9 March 2011 laying down harmonised conditions for the marketing of construction products and repealing Council Directive 89/106/EEC. *Official Journal of the European Union*, L 88: 5–43.
- European Parliament and Council. 2017. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. *Official Journal of the European Union*, L 117: 1–175.
- European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689: 1–144.
- Fernández-Loría, C.; and Provost, F. 2022. Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*, 1(1): 4–16.
- Fischer-Abaigar, U.; Aiken, E.; Kern, C.; and Perdomo, J. C. 2026. Empirically Understanding the Value of Prediction in Allocation. arXiv:2602.08786.
- Fischer-Abaigar, U.; Kern, C.; and Perdomo, J. C. 2025. The Value of Prediction in Identifying the Worst-Off. In *International Conference on Machine Learning*, 17239–17261. PMLR.
- Freiesleben, T.; and Grote, T. 2023. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4): 109.
- Frohnappfel, K.; Seyfert, M.; Bordt, S.; von Luxburg, U.; and Meding, K. 2026. Using predictive multiplicity to measure individual performance within the AI Act. arXiv:2602.11944.
- Fuller, L. L. 1949. The Case of the Speluncean Explorers. *Harvard Law Review*, 62(4): 616–645.
- GENORMA. 2026. prEN 18229-4, AI trustworthiness framework – Part 4: Accuracy. Technical report, CEN/CEN-ELEC.
- Gupta, A.; and Lanteigne, C. 2020. Response by the Montreal AI Ethics Institute to the European Commission’s Whitepaper on AI.
- Hacking, I. 1992. The self-vindication of the laboratory sciences. In *Science as practice and culture*, 29–64. University of Chicago Press.
- Hardt, M. 2025. The Emerging Science of Machine Learning Benchmarks. <https://mlbenchmarks.org>. Manuscript.
- Hébert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 1939–1948. PMLR.
- Hildebrandt, M. 2020. *Law for Computer Scientists and Other Folk*. Oxford University Press.
- HLEG. 2019. Ethics Guidelines for Trustworthy AI. Technical report, European Commission.
- Hu, L. 2026. What is new, and what is old, in fairness and machine learning. *ACM Journal on Responsible Computing*, 3(1): 1–16.
- Hughes, G. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1): 55–63.
- Huntington-Klein, N. 2021. *The effect: An introduction to research design and causality*. Chapman and Hall/CRC, 1st edition.
- Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1859–1876.
- Imbens, G.; and Athey, S. 2021. Breiman’s two cultures: A perspective from econometrics. *Observational Studies*, 7(1): 127–133.
- ISO. n.a. ISO/IEC 25010. Technical report, ISO.
- Jasanoff, S. 2004. Ordering knowledge, ordering society. In *States of knowledge*, 13–45. Routledge.
- Jasanoff, S.; et al. 2004. *States of knowledge*. Taylor & Francis Abingdon, UK.
- Johnson, R. A.; and Zhang, S. 2022. What is the bureaucratic counterfactual? Categorical versus algorithmic prioritization in US social policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1671–1682.
- Kaminski, M. E. 2021. The right to explanation, explained. In *Research handbook on information law and governance*, 278–299. Edward Elgar Publishing.
- Kang, J.; and Buchner, B. 2004. Privacy in Atlantis. *Harvard Journal of Law and Technology*, 18(1): 229–267.

- Koch, B.; and Imai, K. 2025. Statistical decision theory with counterfactual loss. arXiv:2505.08908.
- Koivisto, I.; Koulu, R.; and Larsson, S. 2024. User accounts: How technological concepts permeate public law through the EU's AI Act. *Maastricht Journal of European and Comparative Law*, 31(3): 412–432.
- Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2015. Interpretable classifiers using rules and Bayesian Analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3): 1350–1371.
- Loi, M.; and Heitz, C. 2022. Is calibration a fairness requirement? An argument from the point of view of moral philosophy and decision theory. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, 2026–2034.
- Luhmann, N. 1992. *Die Wissenschaft der Gesellschaft*. Suhrkamp.
- MacCormick, N.; and Weinberger, O. 1986. *An institutional theory of law: new approaches to legal positivism*, volume 3. Springer Science & Business Media.
- Marin, L. G.; Rijsbosch, B.; Meding, K.; Spanakis, G.; van Dijk, G.; and Kollnig, K. 2026. Is your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act. arXiv:2604.03254.
- Meding, K.; and Hagendorff, T. 2024. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, 37(1): 4.
- Möllers, T. M. 2022. *Legal methods: how to work with legal arguments*. CH Beck.
- Molnar, C.; König, G.; Herbringer, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C. A.; Casalicchio, G.; Grosse-Wentrup, M.; and Bischl, B. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International workshop on extending explainable AI beyond deep models and classifiers*, 39–68. Springer.
- Nolte, H.; Rateike, M.; and Finck, M. 2025. Robustness and cybersecurity in the eu artificial intelligence act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 283–295.
- Orr, W.; and Kang, E. B. 2024. AI as a sport: On the competitive epistemologies of benchmarking. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 1875–1884.
- Pearl, J. 2021. Causally Colored Reflections on Leo Breiman's "Statistical Modeling: The Two Cultures" (2001). *Observational Studies*, 7(1): 187–190.
- Peterson, W.; Birdsall, T.; and Fox, W. 1954. The theory of signal detectability. *Transactions of the IRE professional group on information theory*, 4(4): 171–212.
- Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5684–5693.
- Pouget, H.; Holtman, K.; and Emborg, T. 2025. Standard Setting. <https://artificialintelligenceact.eu/standard-setting-overview/>. Accessed: 2026-05-21.
- Recht, B. 2025. The Actuary's Final Word on Algorithmic Decision Making. arXiv:2509.04546.
- Reyes-Cruz, G.; Spors, V.; Muller, M.; Cioffi Felice, M.; Bardzell, S.; Williams, R. M.; Hansson, K.; and Feldfeber, I. 2025. Resisting AI Solutionism: Where Do We Go From Here? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–6.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.
- Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
- Semenova, L.; Rudin, C.; and Parr, R. 2022. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858.
- Simson, J.; Pfisterer, F.; and Kern, C. 2024. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1305–1320.
- Smuha, N.; and Yeung, K. 2025. The European Union's AI Act: Beyond Motherhood and Apple Pie? In *The Cambridge Handbook of the Law, Ethics, Policy and Artificial Intelligence*, 228–258. Cambridge University Press.
- Star, S. L. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, technology, & human values*, 35(5): 601–617.
- Star, S. L.; and Griesemer, J. R. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social studies of science*, 19(3): 387–420.
- Tal, E. 2020. Measurement in Science. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition.
- Vaciago, G. 2026. Accuracy, Robustness and Cybersecurity (Article 15). In *The EU Artificial Intelligence Act: A Thematic Commentary*. Hart Publishing.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887.
- Wang, A.; Kapoor, S.; Barocas, S.; and Narayanan, A. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing*, 1(1): 1–45.
- Weerts, H.; Xenidis, R.; Tarissan, F.; Olsen, H. P.; and Pechenizkiy, M. 2023. Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 805–816.
- Widder, D. G.; and Hicks, M. 2024. Watching the Generative AI Hype Bubble Deflate. *Political Economy of AI Essay Collection*.

- Williams, R. 2014. *Keywords: A vocabulary of culture and society*. Oxford university press.
- Williamson, R. C.; and Menon, A. 2019. Fairness risk measures. In *International Conference on Machine Learning*, volume 97, 6786–6797. PMLR.
- Xu, R.; Wang, Z.; Fan, R.-Z.; and Liu, P. 2024. Benchmarking benchmark leakage in large language models. arXiv:2404.18824.
- Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.