LLMs, legal contestability and scientific falsifiability

Mireille Hildebrandt[*]

Abstract

In this chapter Hildebrandt investigates the link between the contestability that is key to constitutional democracies on the one hand and the falsifiability of scientific theories on the other hand, with regard to large language models (LLMs). Legally relevant decision-making that is based on the deployment of applications that involve LLMs must be contestable in a court of law. The current flavour of such contestability is focused on transparency, usually framed in terms of the explainability of the model (explainable AI). In the long run, however, the fairness and reliability of these models should be tested in a more scientific manner, based on the falsifiability of the theoretical framework that should underpin the model. This requires that researchers in the domain of LLMs learn to *abduct* theoretical frameworks, based on the output models of LLMs and the real world patterns they imply, while this abduction should be such that the theory can be inductively tested in a way that allows for falsification. On top of that, researchers need to conduct empirical research to enable such inductive testing. The chapter thus argues that the contestability required under the Rule of Law, should move beyond explanations of how the model generates its output, to whether the real world patterns represented in the model output can falsify the theoretical framework that should inform the model.

Keywords

Rule of Law, contestability, LLMs, falsifiability, theory, inductive fallacy, abduction, explainable AI

[*] Mireille Hildebrandt is Emerita Professor of Law at Vrije Universiteit Brussel (Faculty of Law and Criminology) and at Radboud University (Science Faculty).

1 Introduction

In her seminal *Science at the Bar*,[1] Sheila Jasanoff demonstrated how the contestability that is key to the adversarial trial turned out to be key to scientific practice. Notably with regard to DNA fingerprinting, '[i]n an effective display of boundary work, DNA fingerprinting was originally represented by its proponents as a taken-for-granted technique belonging only to the fields of molecular genetics and molecular biology'. Due to adversarial expert testimony in a range of court cases, culminating in the O.J. Simpson case, it turned out that without integrating population genetics DNA fingerprinting was unreliable and in point of fact *biased against black defendants*. Perhaps it is time to test the reliability of so-called AI models, deployed in a whole range of sectors and applications, based on appropriate scientific methodology, if needs be in a court of law.

This chapter highlights the link between legal contestability and scientific falsifiability, notably in relation to the development and deployment of AI models. Many of the problems that result from the deployment and use of these models concern safety, health and/or infringements of fundamental rights. In a legal context it is crucial that the output of AI models can be contested in a court of law, due to its impact on individual safety, health and fundamental rights, and on public goods such as critical infrastructure, climate change, employment and insurance. What matters here is the fairness and the reliability of AI models and the ability to check both. Over the past decades, these issues have been defined in terms of the transparency, explainability and interpretability of specific AI applications, often focused on individual decisions or behaviours. The EU General Data Protection Regulation (GDPR) attributes an individual right to data subjects to obtain 'meaningful information about the logic involved', 'the significance and the envisaged consequences' of automated decision-making (art. 15.1(h) jo art. 22 GDPR). This right (and the complementary legal obligation for data controllers to provide such meaningful information in art. 13 and 14 GDPR) has triggered the rise of a subdomain in computer science and engineering, under the heading of explainable AI. This way computer scientists contribute to making such decision-making contestable, as required in recital 71 GDPR: 'the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision').

Recital 71 GDPR indeed centres its attention on the contestability of individual decision-making, thus highlighting the need for contestable decisions:

> In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, (…).

Similar obligations are imposed on providers of high risk AI systems in the EU AI Act, focusing on the reliability and fairness of these systems, for instance requiring that they enable 'the natural persons to whom human oversight is assigned' by the deployer, to 'understand the relevant capacities and limitations of the high risk AI system' and to 'interpret its output' (art. 14 AI Act), while also imposing e.g. a whole range of data governance obligations (art. 10 AI Act) that should allow others to assess the reliability and fairness of high risk AI systems. Though here the focus is not on individual decisions or

[1] Sheila Jasanoff, *Science at the Bar. Law, Science, and Technology in America* (Harvard UP 1995)., p. 56-57.

behaviours but on the foreseeable risk of deploying a high risk AI system for its intended purpose or for other foreseeable purposes. There is no attempt to dig deeper into the scientific grounding of the underlying AI models and their implied theoretical underpinnings.

In this chapter, I propose to open a new research agenda in the domain of AI research, building on the distinction made by Popper and Peirce between scientific and other types of knowledge. I will argue that data-driven AI models often fall prey to the inductive fallacy, making their output fundamentally unreliable, noting that the evaluation and testing of AI models is notoriously challenging. Following Peirce, scientific theories should aim to explain basic empirical findings based on abduction, in such a way that the theory can be inductively tested in a way that could falsify the theory (not the model). At this moment testing and evaluation boils down to inductive testing to check inductive inferences, lacking a proper theoretical framework that could offer a falsifiable explanation of the real-world patterns supposedly detected in the relevant training data. It would be very interesting to see the reliability of LLMs challenged in court, based on the contention that the claims made on their behalf are not corroborated by genuinely scientific evidence, thus challenging the methodological integrity of much privately funded AI research.

After explaining the key tenet of contestability of legally relevant decision-making in constitutional democracies in section 2, I will discuss falsification and fallibilism in science in section 3, followed by the implications thereof for the legal contestability of AI models in section 4. Conclusions will be drawn in in section 5.

2 Contestability in law

The core tenet of the Rule of Law is the contestability of legally relevant decision-making, in an independent court of law. The Rule of Law thus institutionalises the idea that those who have an interest in taking or justifying a specific decision do not get to decide on its legality or lawfulness and builds on the idea that adversarial or contradictory procedures are crucial to both fair and to reliable decision making in law.

Safeguarding contestability is only possible if the Rule of Law is not about a sovereign who is willing to bind themselves to the rules they enact, because self-binding would make their subjects dependent on the sovereign's good intentions and reliable insights. In the end, self-binding results in an arbitrary rule.[2] Instead, the Rule of Law is about dedicated institutional settings that ensure countervailing powers against the authority of the state (while limiting the power of other big players).[3] In other work I have explained this by referring to Odysseus' and the Sirens,[4] who were known to seduce travellers with their songs, such that they could not leave, resulting in death of starvation. To prevent being lured into their enchantments, while nevertheless hearing their beguiling songs, Odysseus did two things. He tied himself to the mast of the ship that would sail through the land of the Sirens. But he knew that would

[2] Gerald J Postema, 'The Idea of the Rule of Law' in Gerald J Postema (ed), *Law's Rule: The Nature, Value, and Viability of the Rule of Law* (Oxford University Press 2023) <https://doi.org/10.1093/oso/9780190645342.003.0001> accessed 18 August 2025.

[3] Koen Lenaerts, 'On Checks and Balances: The Rule of Law within the EU European Union's Horizon 2020 Research and Innovation Programme' (2023) 29 Columbia Journal of European Law 25 <https://heinonline.org/HOL/P?h=hein.journals/coljeul29&i=182> accessed 18 August 2025.

[4] Mireille Hildebrandt, *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology* (Edward Elgar 2015).

not be enough, as he would inevitably call on his sailors to untie him as soon as he heard the Siron songs. Therefore, before tying himself to the mast, he waxed the ears of his sailors, to prevent them from following his pleas to untie him. This shows why self-binding is not enough. In terms of constitutional law, self-binding would protect a rule *by* law, rather than a rule *of* law. Montesquieu's ingenious *iudex (mh: non rex), est lex loquens* makes precisely this point: the judge, not the executive or the legislature, speaks the law.[5]

Independence is meant to afford fair and reliable decision making, as an affordance of the contestability it enables. In law contestability refers to justice, legal certainty and the instrumentality of the law.[6] Justice concerns fairness, that is treating equal cases equal and unequal cases unequally to the extent of their inequality. Fairness, thus, requires discernment as to the relevance of different circumstances to decide which cases are equal or to what extent they are not equal. Such discernment requires judgement rather than calculation,[7] as the latter must assume and thus hide the qualification of the facts in terms of a legal norm and the choice and interpretation of the relevant legal norm(s) in light of the facts. Legal certainty should enable those who share jurisdiction to foresee the consequences of their actions. Legal certainty, thus, concerns the trust that the law will be enforced after being 'posited' by the legislature, based on the interpretation of the courts, which requires a delicate balance between foreseeability and contestability, highlighting the argumentative nature of positive law.[8] The instrumentality of law concerns the purpose of the law. Instrumentality, meanwhile, should not be confused with instrumentalism, where law is but a neutral tool to achieve policy goals. Instrumentality, instead, refers to the law as an instrument to achieve policy goals determined by the legislature, while requiring that the values and goals of the law are taken into account. The latter means that the law can never be a neutral instrument (noting that both justice and legal certainty must be safeguarded).[9]

Though some may believe that the contestability of legally relevant decision-making is a matter of justice and fairness, it should be clear that it is also about reliability. In law, such reliability concerns (1) the law of evidence (establishment of the facts), (2) the interplay between justice (the decision on what cases are similar and must thus be treated the same) and legal certainty (the foreseeability of the legal consequences of one's actions based on both enforcement and equal treatment) and (3) the purposiveness or instrumentality of the law, because the interpretation of a legal text is informed by the purpose of the law, which in turn directly relates to justice and legal certainty.

Though some may believe that the role of contestability in science is only relevant in relation to the law of evidence, highlighting the specificity of the legal domain compared to the domain of science, I will argue, that also in science, reliability is directly linked with contestability and with our ability to foresee the consequences of our actions. Such reliability plays out in an existential way in the case of climate change, allowing us to distinguish between the contestability of climate science on the one hand and

[5] KM Schoenfeld, 'Rex, Lex et Judex: Montesquieu and La Bouche de La Loi Revisted' (2008) 4 European Constitutional Law Review 274.

[6] Mireille Hildebrandt, 'Radbruch's Rechtsstaat and Schmitt's Legal Order: Legalism, Legality, and the Institution of Law' (2015) 2 Critical Analysis of Law <http://cal.library.utoronto.ca/index.php/cal/article/view/22514> accessed 24 March 2015; *The Legal Philosophies of Lask, Radbruch, and Dabin* (Harvard University Press 2014) <http://www.degruyter.com/view/product/252229> accessed 27 May 2014.

[7] Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (The MIT Press 2019).

[8] JEREMY WALDRON, 'THE RULE OF LAW AND THE IMPORTANCE OF PROCEDURE' (2011) 50 Nomos 3 <https://www.jstor.org/stable/24220105> accessed 17 January 2020; *The Legal Philosophies of Lask, Radbruch, and Dabin* (n 7).

[9] Ronald Dworkin, *Law's Empire* (Fontana 1991).

the denial of climate change on the other. For the same reason, contestability and reliability are also highly relevant for the development and deployment of AI models and systems.

3 Falsification and fallibilism in science

3.1 Popper's falsification

In science, Popper advocated falsification as the only reliable method to test the scientific validity of a specific scientific theory or hypothesis that is meant to explain observable phenomena. For Popper, falsifiability was the criterion that distinguishes science from other types of knowledge.[10]

Falsifiability means that a hypothesis or theory that aims to explain observable phenomena must be articulated in such a way that it allows for the deduction of hypotheses that can be tested against basic empirical facts. Instead of looking for verification, Popper asserted that the scientific quality of a theory depends on its falsifiability; researchers should not look for the confirmation of their theories but for phenomena that may contradict them. If a theory on the genetic make-up of swans implies that all swans are white, we need to look for black swans, not for white ones. Discovering a black swan will invite either the rejection of the entire theory or its amendment and refinement in a way that explains the occurrence of black swans. This example should also clarify that 'all swans are white' is not a theory, but a hypothesis that would enable the testing of a theory. Science is not about mere extrapolation (all swans I saw were white, so all swans are white) but about theoretical frameworks that can explain the observation that all swans are white. That explanation can be refuted by the perception of a black swan.

Interestingly, Popper's emphasis on basic empirical facts is coupled with his insistence that perception of such facts is 'theory laden'. In fact, Popper does not speak of 'basic empirical facts' but of 'basic empirical statements', acknowledging that even these can be refuted. We cannot observe the world without background knowledge that allows us to single out certain observations as relevant. In a sense, Popper thus already integrated Kuhn's insights regarding the role of paradigms in scientific research, even though paradigms do not equate with scientific theory. One could say that Kuhn's paradigms do for theories what Popper's theories do for observable facts; they both admit that the core building blocks of scientific enquiry depend on frameworks that function as mostly invisible vanishing points, though reality may at some point force us to address these vanishing points (whether theories or paradigms). Contestability and notably falsifiability, therefore, are of key importance for individual and collective survival and flourishing, because getting the theory wrong or working within the context of a problematic paradigm has real-world and real life consequences.

The first edition of Latour's and Woolgar's *Laboratory Life* had the subtitle 'The social construction of scientific facts'.[11] From the second edition onwards the word 'social' was removed.[12] According to Latour, scientific findings are not a matter of merely 'social' construction, but a complex experimental

[10] Karl Popper, *The Logic of Scientific Discovery* (2 edition, Routledge 2002); Stephen Thornton, 'Karl Popper' in Edward N Zalta and Uri Nodelman (eds), *The Stanford Encyclopedia of Philosophy* (Winter 2023, Metaphysics Research Lab, Stanford University 2023) <https://plato.stanford.edu/archives/win2023/entries/popper/> accessed 18 August 2025.

[11] Bruno Latour and Steve Woolgar, *Laboratory Life: The Social Construction of Scientific Facts* (SAGE Publications, Inc 1981).

[12] Bruno Latour, *Laboratory Life: The Construction of Scientific Facts by Latour, Bruno, Woolgar, Steve Published by Princeton University Press* (Princeton University Press).

fabrication that aims to invite and address a reality that offers resistance against wrong interpretations.[13] Latour's brilliant observations in the laboratory of Roger Guillemin (who later won the Nobel Prize) asserted that scientific findings are always also social, but what matters is how experimental laboratory science lures 'reality' into answering specific scientific questions. What Popper and Kuhn and Latour highlight is that in scientific research, this grappling with reality is deeply contingent upon complex theoretical frameworks that in-form our perception and understanding of what may seem objective facts, noting that these frameworks may operate as tacit knowledge or common sense, preventing us from taking another perspective that would reveal other facts that require new lenses to be made visible. Falsification then, becomes pivotal, though not obvious, to loosen the grip of these frameworks.

For example, in her brilliant *Turning to Stone*,[14] Bjornerud describes the turbulent history of the science of geology, including the hold of a succession of problematic theories and paradigms on the research done in geology. Geology is not based on scientific experiments in a laboratory, though in the course of its existence as a science a great number of other disciplines became involved that were indeed grounded in experimental sciences (for instance minerology, petrology, cosmochemistry and nuclear physics to aid geochronology). In her book *Timefulness*,[15] she describes the nefarious impact of the physical sciences with their foundational assumptions of timelessness on the understanding of 'deep time' (the history of the earth before the emergence of living organisms) that is crucial for our understanding of the future of planet earth. The timelessness that pervades the physical sciences is key to a mathematical, neoplatonic understanding of reality,[16] which – unsurprisingly - similarly prevails in computer science and is therefore highly relevant for the contestability of AI. The development of geology as a science, demonstrates both the important role of falsification by unruly facts that contradict prevailing theories and the difficulty to agree on the facts, depending on one's theoretical assumptions.

3.2 Peirce's fallibilism

Charles Saunders Peirce, the founder of US philosophical pragmatism, who was also a semiotician and a scientist, had already advocated fallibilism as the test case of the scientific method.[17] Fallibilism refers to the position that no knowledge or belief can be justified in a final way, and that scientific knowledge can always be refuted based on new information.[18] In that sense Popper was also a fallibilist, arriving at the same conclusions as Peirce with regard to the nature of scientific knowledge.

What makes Peirce even more interesting is the way he described scientific research. According to Peirce a scientist (1) develops a theory by way of abduction, that is by developing a potential explanation of a specific set of phenomena, (2) derives hypotheses from such a theory by way of

[13] See also more recently Kofman, 'Bruno Latour, the Post-Truth Philosopher, Mounts a Defense of Science' *The New York Times* (25 October 2018) <https://www.nytimes.com/2018/10/25/magazine/bruno-latour-post-truth-philosopher-science.html> accessed 18 June 2022.

[14] Marcia Bjornerud, *Turning to Stone: Discovering the Subtle Wisdom of Rocks* (Flatiron Books 2024).

[15] Marcia Bjornerud, *Timefulness: How Thinking Like a Geologist Can Help Save the World* (Princeton University Press 2020).

[16] Dan McQuillan, 'Data Science as Machinic Neoplatonism' [2017] Philosophy & Technology 1 <https://link.springer.com/article/10.1007/s13347-017-0273-3> accessed 30 August 2017.

[17] Charles Saunders Peirce, *Selected Writings, Edited with an Introduction and Notes by Philip P. Wiener* (Dover 1958). Chrysogonus M Okwenna, 'Peirce's Fallibilism: A Thematic Analysis and the Revisitation of the Origins of Fallibilism' (2021) 19 Amamihe: Journal of Applied Philosophy 18.

[18] 'Fallibilism | Internet Encyclopedia of Philosophy' <https://iep.utm.edu/fallibil/> accessed 18 August 2025.

deduction (logic), (3) follows this up with testing whether the hypotheses can be disproven by way of induction (facts). The key point of both Popper and Peirce was that mere verification does not offer much help, as it induces confirmation bias. The specific point of Peirce was that the inductive method is about testing, not about inference (which would require abduction). Inductive inferencing, is liable to the inductive fallacy that asserts that verification is not a proper test for scientific theories; the fact that the sun rises everyday does not in itself guarantee that it will rise tomorrow morning.[19] This point has major relevance for data-driven approaches, which thrive on inductive inference and are thus necessarily liable to the inductive fallacy.

Abduction signifies a jump from observations to their explanation, it requires creativity and an open mind, because abduction is not a matter of either deductive or inductive logic. Different explanatory theories are possible and their construction is not based on a method or algorithm but on relevant knowledge and intuition, relevant experience, daring and foresight. Inductive inference is prone the inductive fallacy that assumes that past patterns will be repeated in the future. Though this may be true, it is not necessarily the case and does little to explain how the pattern came about, which would provide a better understanding of the future.

4 Fallibilism, theory and AI models

4.1 AI models and the philosophy of science

In this chapter, I focus on so-called large language models (LLMs) that have been hailed as a revolution in search, text-generation, medicine, employment, legal practice and finance. LLMs are based on a specific type of machine learning, usually referred to as deep learning, where algorithms are pretrained on Big Data, to uncover hidden correlations, also called distributions. In the layered and complex process of uncovering such distributions, the model is able to generate text, based on calculating a sequence of predictions of next-tokens (word parts), given a prompt. The sequence of predictions of next-tokens is basically a prediction of language behaviours, based on massive amounts of language-behaviour data. As data scientists (should) know, machine learning is based on the incorrect but useful assumption that the distribution of the training data is equal to, or at least similar to, the distribution of future data. We cannot, however, take that for granted, and as we cannot train a model on future data, the problem will persist until we acknowledge the limits of inductive inferences in AI.[20] This makes Popper's and Peirce's insistence on falsification and fallibilism highly relevant for the testing of the implied theoretical assumptions of AI models.

Developing computer code is not necessarily a scientific undertaking and software developers are not necessarily scientists or even interested in science. However, when presenting an AI model as a business proposition, to invite funding, it should be reliable in the straightforward sense that when implemented in real-world scenarios it will offer the claimed functionality. Such reliability is directly related to safety, possibly to health and may also impact on the substance of human rights, notably – though not only - non-discrimination and privacy. All depending on the domain of application and the claimed functionality, e.g. producing specified briefs or records in the correct format; answering questions about health, political choices, architecture, art or education; generating the molecular structure of existing

[19] Leah Henderson, 'The Problem of Induction' in Edward N Zalta and Uri Nodelman (eds), *The Stanford Encyclopedia of Philosophy* (Winter 2024, Metaphysics Research Lab, Stanford University 2024) <https://plato.stanford.edu/archives/win2024/entries/induction-problem/> accessed 31 August 2025.

[20] See e.g. Zhicheng Lin, 'Six Fallacies in Substituting Large Language Models for Human Participants' (2025) 8 Advances in Methods and Practices in Psychological Science 25152459251357566 <https://doi.org/10.1177/25152459251357566> accessed 18 August 2025.

and/or potentially unknown or even new proteins; explaining how to build specific types of weapons; generating relevant computer code to solve specified problems etc. Due to the real-world implications of AI models, it makes sense to investigate to what extent the methodologies that underlie the development these models allow for falsification rather than verification and how this relates to (1) the use of so-called performance metrics in the context of machine learning (as they seem focused on verification), (2) adversarial machine learning (focused on falsifying a specific model or system) and (3) the a-theoretical nature of much research in machine learning models or systems (basically succumbing to the inductive fallacy). In other words, to what extent does the testing of AI models align with contestation or falsification rather than verification.

4.2 Performance metrics

In machine learning, three performance metrics have been put forward to testify to the reliability of an ML system: accuracy, precision and recall.[21] They present how often the model gets it right, in terms of false positives (FPs), false negatives (FNs), true positives (TPs) and true negatives (TNs). Precision checks a model's TPs in relation to all the model's positives (TPs plus the FPs). Recall checks a model's TPs in relation to all positives (TPs plus FNs). Accuracy checks a model's TPs and TNs in relation to all models outputs (TPs plus TNs plus FPs plus FNs).

These metrics are interesting but their relevance depends on both the domain of application (do we need to exclude FNs or rather FPs) and on the distribution of the data (if there are few positives in the data, different metrics will offer very different test results). For instance, if out of 100 testcases of breast cancer only 4 people actually have cancer (positives), while all other 96 people do not have cancer (negatives), and the model correctly predicts cancer for 3 out of the 4 positives, while also incorrectly predicting cancer in 7 other cases, the accuracy will be 92%, the precision will be 75% and recall will be 30 %. The latter predicts the chance that a person who is diagnosed with cancer has a 30% chance that the diagnosis is correct. For this type of domain (breast cancer diagnosis) recall seems the better metric, even though the accuracy and the recall may sound far more impressive.

The role of falsification is unclear, because we are dealing with statistical inferences. Determining that the model gets it wrong (a FP or a FN) does not invalidate the model, it rather offers a way to evaluate its reliability. Detecting FPs or FNs does not falsify the model, because the model is neither a scientific theory nor a scientific hypothesis. Both Popper and Peirce admitted that in the case of probabilistic theories, falsification becomes problematic. If my theory is that most swans are white, a black swan does not falsify. In order to falsify this 'theory' we would need to quantify 'most'. However, 'most swans are white' is not really a theory, because it explains nothing, it merely posits an empirical fact. In section 0 I will discuss the role of theory in AI, explaining that LLMs are models but not theories. They produce outputs that can be evaluated, but they do not offer any explanation that can be tested by way of falsification.

Accuracy, recall and precision are not obvious for the testing of LLMs. As a generative technology, pretrained on massive amounts of language-behaviour data (which is often not disclosed), the distinction between TP, FP, TN and FN is not very clear. Developers of LLMs sometimes claim that their models' output does not qualify as true or false, because they don't do anything other than build on complex stochastic patterns in our past language-behaviour data. This implies that one could fact-check the output against real-world phenomena,[22] but as we all know, these models provide different

[21] Sathyanarayanan Swaminathan and B Roopashri Tantri, 'Confusion Matrix-Based Performance Evaluation Metrics' (2024) 27 African Journal of Biomedical Research 4023.

[22] Florian Le Bronnec and others, 'Exploring Precision and Recall to Assess the Quality and Diversity of LLMs' (arXiv, 4 June 2024) <http://arxiv.org/abs/2402.10693> accessed 21 September 2025.

answers with each iteration and it is not possible to predict which answer they will give, as this also depends on the kind of prompts users deploy to query the model. On top of that, developers claim that by pre-prompting the model before releasing it to potential users, they can in principle align the model with 'human values'. This, however, raises the question of whose values must be aligned with and who gets to determine what values must be incorporated, noting that after releasing the model, groups of users (including bots) could work to align the model with other kinds of values (called prompt injections, see section 4.3 below).

Narayanan and Kapoor have indicated how and why the evaluation of LLMs may run into trouble due to prompt sensitivity, construct validity or contamination.[23] If the output of an LLM depends on what prompts are given it becomes difficult to determine anything like accuracy, precision or recall, as the output will keep fluctuating (they qualify this kind of prompt sensitivity as a reproducibility problem). If the output of the LLM is meant to align with a construct (targeted output) that is actually co-defined by the users, then to test the accuracy, precision or recall we would need access to the prompt distribution (which is not usually provided, either due to trade secrets, IP rights or privacy issues). This means that the validity of the construct cannot be tested. The third problem they point out concerns contamination, which refers to the issue of whether these models use memorisation to 'get things right', which may result in flipping from 100% accuracy to 0% once the benchmark that is memorised is no longer valid. Contamination can also be caused by so-called leakage, where the input contains hidden information the model is expected to predict.

In a sense, the trouble with testing LLMs can be situated in the feedback loops between training data (which are by definition historical data), prompts (whose distribution is not shared) and output. The *targeted* output may be hard to determine if the actual output has a performative effect, in the sense that given prompt sensitivity and contamination, we cannot determine the real-world value of the targeted output and may deploy the actual output as if it were the targeted output.

Whether or not we understand the technical intricacies of these types of metrics in the case of LLMs, it should be clear that falsification has a long way to go and would require a combination of access to both pretraining data and prompt distribution and actual empirical research to provide us with findings about the targeted output, which would have to enable basic empirical statements to count as falsification. Noting also that it is unclear to what extent an LLM would qualify as the type of theory or hypothesis that Popper and Peirce have in mind, that is a theory or hypothesis that explains real-world phenomena, which I will discuss below under section 0.

4.3 Adversarial machine learning

Adversarial machine learning has been understood in two ways. First, one type of adversarial machine learning, which I will refer to under the abbreviation AML, concerns the security of ML and AI systems. AML investigates their vulnerability to AML attacks that disrupt the expected behaviour of an ML or AI system, by way of data poisoning, prompt injection or other adversarial interventions that result in unforeseen unreliable output. The term thus refers to both the attacks and their anticipation in order to prevent them.[24] Resilience against such attacks would qualify systems as 'robust', based on effective

[23] See the powerpoint presentation of Arvind Narayanan & Sayash Kapoor, *Evaluating LLMs is a minefield*, Oct 4, 2023 https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/#/6, with a number of relevant references to their work in this domain, e.g. Sayash Kapoor and others, 'AI Agents That Matter' (arXiv, 1 July 2024) <http://arxiv.org/abs/2407.01502> accessed 17 September 2025. And the Holistic Agent Leaderboard (HAL), a webtool that allows the testing and evaluation of AI agents: https://hal.cs.princeton.edu.

[24] Jasmita Malik, Raja Muthalagu and Pranav M Pawar, 'A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies' (2024) 12 IEEE Access 99382 <https://ieeexplore.ieee.org/document/10584534/> accessed 3 September 2025; Apostol Vassilev and others, 'Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations' [2025] NIST

threat modelling. Such resilience will be key in healthcare, semi-autonomous transportation, insurance, financial systems, social security, immigration and so on. Especially with regard to generative AI, prompt injections enable the manipulation of generative AI systems, noting that due to their stochastic nature these systems are unreliable by definition, making it even harder to figure out how prompt injections made a difference. This type of AML could be termed as a kind of falsification, because, like other types of penetration testing, AML seeks to uncover and exploit vulnerablities, or to prevent their exploitation. AML is not, however, focused on falsifying the underlying theory or explanation of these systems. In that sense, AML must not be understood as if it presents us with Popper's falsifiability, which is focused on testing a theory (that is, an explanation of empirical findings). Nevertheless, AML shows us the vulnerabilities of systems that embody statistical patterns without offering any kind of overarching theory. The latter would help make these systems more robust against future breakdowns, because a scientific theory would integrate the contextual and temporal conditions that explain the behaviour of the system, enabling developers to foresee future dependencies.

Another use of the term adversarial machine learning would align with Disalvo's Adversarial Design[25] and my own 'Agonistic Machine Learning',[26] where in both cases, the idea is to develop a plurality of models and/or systems, based e.g. on different training data, different prompts, different ML techniques, in order to play around with them, test them against each other and uncover hidden assumptions that could be said to represent the tacit theoretical underpinnings of the model or system. By observing or foreseeing the real-world implications of a plurality of models or systems that aim to 'solve' the same real-world problem with different ML designs, their underlying theoretical beliefs may be uncovered and tested. Agonism is an approach that builds on Mouffe's participatory democratic theory,[27] and on Rip's constructive technology assessment.[28] In both cases, the idea is that robustness derives from the confrontation of alternative ideas and approaches, requiring an agonistic debate that brings out a plurality of different arguments, allowing those who develop public policies and/or technological innovations to foresee potential functionalities that contribute to public and individual wellbeing as well as potential risks to health, safety and fundamental rights. Though agonistic debates in the realm of political decision-making and technical transformation do not necessarily present us with the falsification of scientific theories, they may nevertheless uncover myriad assumptions that are taken for

<https://www.nist.gov/publications/adversarial-machine-learning-taxonomy-and-terminology-attacks-and-mitigations-0> accessed 3 September 2025; 'How AI Can Be Hacked with Prompt Injection: NIST Report | IBM' (19 March 2024) <https://www.ibm.com/think/insights/ai-prompt-injection-nist-report> accessed 3 September 2025; Matt Sutton and Damian Ruck, 'Indirect Prompt Injection: Generative AI's Greatest Security Flaw' (Centre for Emerging Technology and Security The Alan Turing Institute 2024) Expert Analysis <https://cetas.turing.ac.uk/publications/indirect-prompt-injection-generative-ais-greatest-security-flaw> accessed 17 September 2025.

[25] Carl DiSalvo, *Adversarial Design* (Reprint edition, The MIT Press 2015).

[26] Mireille Hildebrandt, 'Privacy As Protection of the Incomputable Self: Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law <https://papers.ssrn.com/abstract=3081776> accessed 12 December 2017.

[27] Chantal Mouffe, *The Democractic Paradox* (Verso 2000); CHANTAL MOUFFE, 'Deliberative Democracy or Agonistic Pluralism?' (1999) 66 Social Research 745 <http://www.jstor.org/stable/40971349> accessed 28 November 2017.

[28] Arie Rip, Thomas Misa and Johan Schot, *Managing Technology in Society: The Approach of Constructive Technology Assessment* (Pinter Publishers 1995); Arie Rip, 'Science for the 21st Century', *The Future of the Sciences and Humanities. Four analytical essays and a critical debate on the future of scholastic endeavour* (Amsterdam University Press 2002) <https://research.utwente.nl/en/publications/science-for-the-21st-century> accessed 3 September 2025. at 117, also referring to Popper's fallibilism.

granted in the domains of political theory, computer science and law, potentially making those assumptions and their theoretical underpinnings falsifiable.

4.4 The role of theory in the development of LLMs

By way of example, we can think of the following. Imagine that we use generative AI for search, instead of the 'traditional' search engines and we largely rely on its output. Agonistic machine learning would imply that while building LLMs, developers and potential deployers as well as those targeted by the model, will engage in an iterant discussion about the functionality of the model, while probing into the assumptions that must be made to build an LLM. As we saw above, the most important assumption is that the distribution of pre-training data is similar to that of future data. This assumption falls prey to the inductive fallacy, because we continuously change our language-behaviour to accommodate our changing environments or to deliberately change our environment by introducing new institutional facts (which are created by speech acts).[29] LLMs, therefore, will always lag behind. On top of that, massive amounts of data easily generates myriad spurious correlations that fit the distribution of the training data but do not offer real-world patterns, resulting in hallucinations. The theory underpinning the construction of LLMs seems to include a strong belief in the relevance of inductive inferences and a similarly strong belief that complex mathematical patterns in data are a proxy for relevant real-world patterns.

In principle, such patterns should trigger abductive inferences that can explain real-world observable facts. The problem with current LLM development, is that no effort is made to develop scientific theories, except, perhaps, in niche domains, such as biomedical research (AlphaFold).[30] Once we have such abductive inferences, consisting of theoretical frameworks that can explain real-world patterns supposedly presented by LLMs, we could start testing and falsifying them, by conducting empirical research capable of falsifying the theoretical framework. Such frameworks would be seriously helpful in rejecting or refining models that are falsified, for instance by iterant hallucinations. What have been coined 'hallucinations' refer to patterns in the data that do not stand for patterns in the real-world. At this moment, however, checking the output of an LLM for errors does not equate with falsification, because no theory is put forward. This makes it hard to trust the model, because the only way to rely on its output is to test it against real-world phenomena, fact-checking for relevance, bias, omissions and hallucinations. Such fact-checking would make the model inefficient, as it would take up quite some time and relevant expertise, while skipping the test could make it ineffective because those who deploy it may be building on sand.

[29] JL Austin, *How to Do Things with Words* (2nd edn, Harvard University Press 1975).

[30] Ștefan-Bogdan Marcu, Sabin Tăbîrcă and Mark Tangney, 'An Overview of Alphafold's Breakthrough' (2022) 5 Frontiers in Artificial Intelligence <https://www.frontiersin.org/articles/10.3389/frai.2022.875587> accessed 30 December 2022; Mireille Hildebrandt, 'Ground-Truthing in the European Health Data Space', *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5 HEALTHINF: BIOSTEC* (Scitepress 2023) <https://www.scitepress.org/Link.aspx?doi=10.5220/0011955900003414> accessed 27 March 2023.

In other work,[31] I made a more in-depth analysis of the role of theory in machine learning, quoting Chris Anderson's 2008 *Wired* article on 'The End of Theory':[32]

> At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising – it just assumed that better data, with better analytical tools, would win the day. And Google was right.

Anderson was, however, not right. Mathematics and statistics can be deployed in numerous ways when it comes to search, depending on what the developers of the page-rank algorithms deem relevant. The idea that such relevance would be solely determined by a combination of collective an individualised aggregates of search behaviours and the amount and intensity of hyperlinks between webpages is rather naïve. As Brin and Page, the founders of Google, wrote in their 1998 article on the page-rank algorithm,[33] commercial interest and business models based on behavioural advertising will introduce the interests of the provider of the search engine and/or those of advertising intermediaries and/or those of whoever purchases advertising space. This transforms the relevance of search output. Without a theory on the relevance of search results the search engine cannot be tested, enabling advertising intermediaries (who do the math) to manipulate advertisers into believing that their investments in digital advertising are lucrative, while users can be manipulated into believing that the search results are ranked in a neutral, objective way, offering the best possible information available. Once the providers of search engines usurped the advertising intermediaries the circle was closed and falsification became next to impossible, noting that his would concern falsifying the – undisclosed – search algorithm, not any kind of theory. One could abduct theories, for instance suggesting that search results are ranked to favour specific types of advertisers or other parties, or, ranked to increase 'engagement' with the search engine. However, testing this would require reverse profiling and massive amounts of user data, often resisted by the provider. Note that the EU Digital Services Act includes specific requirements for very large online platforms (VLOPs), meant to enable such reverse profiling to e.g. uncover unfair trade practices. Again, it is unclear at which point such a 'theory' qualifies as a scientific theory in the sense of Popper and Peirce. Nevertheless it seems pivotal that the hidden assumptions that inform search engines and other AI models are mined and researched in a way that enables scientific research, in order to assess the reliability of their output. This would entail a theoretical framework that can be falsified based on relevant observations and this, in turn, requires that such a framework becomes contestable.

As indicated, niche domains such as biomedical research may be more open to abductive inferencing and inductive testing. For instance, AlphaFold, capable of predicting the structure of proteins, including the 'discovery' of unknown and possibly non-existing proteins, could trigger a whole range of theoretical explanations. Together with experimental testing, this could enable promising new scientific insights in both health and disease, though without experimental testing and devoid of theoretical

[31] Hildebrandt, *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology* (n 5). at 25-26 and 37-40.

[32] Chris Anderson, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' (2008) 16 Wired Magazine. At x, quoted in: Hildebrandt, *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology* (n 5). at 25.

[33] Sergey Brin and Lawrence Page, 'The Anatomy of a Large-Scale Hypertextual Web Search Engine' (1998) 30 Computer Networks and ISDN Systems 107 <http://www.sciencedirect.com/science/article/pii/S016975529800110X> accessed 18 October 2012.

exploration it could create unprecedented health hazards, such as the deployment of synthetic proteins whose ecological impact is unknown. A recent article published in *Nature*[34] has presented the results of a dedicated LLM-type of generative AI model, Delphi-2M, trained on 400.000 UK patient data, with regard to 1.9 Danish individuals, whose data the model had not 'seen'. The model was capable to predict the patterns of disease progression of the Danish individuals with an uncanny accuracy. The article notes that '[u]nderstanding each individual's multi-morbidity risks is important to tailor healthcare decisions, motivate lifestyle changes or direct entrance into screening programs, as is the case for cancer', suggesting that these kinds of predictions are of key relevance for both individual healthcare decision-making and for national and global healthcare strategies. Interestingly, the article devotes a section to 'Explaining Delphi-2M predictions', to explain 'how Delphi-2M uses past information to predict future', focusing on the inductive inferences, without any exploration of potential abductive inferences that could explain real-world disease interaction patterns. This should not be a problem, as long as subsequent research into the causes of the relevant dependencies results in meaningful explanations that inform real world healthcare decision-making, both in individual cases and in healthcare policy. As I have explained in previous work, the deployment of this kind of modelling in the context of health may be both beneficial and extremely hazardous, due to the need for experimental testing without which long term health hazards may occur and due to its potential dual use in the context of biological warfare.[35] In this chapter I add the need for abductive inference to develop falsifiable theoretical frameworks, which should contribute to long-term health benefits at the level of public health and prevent the consequences of automation bias and deskilling.

5 Conclusions: AI, fallibilism and contestability

The digital transformation, that is heralded as a core EU policy objective,[36] makes us increasingly dependent on the use of AI models that have no theoretical grounding and whose long-term real-world reliability can therefore not be assessed. In this final section, I will conclude that the contestability of legally relevant decision-making, that is core to constitutional democracies, requires keen attention to the reliability of these systems, asserting that this should involve the ability to falsify their underlying theoretical assumptions.

Such falsification depends, as we have seen above, on two key requirements. First, we need to learn to abduct theoretical frameworks, based on the output models of LLMs, while this abduction should be such that the theory can be inductively tested in a way that allows for falsification. Second, we need to conduct empirical research to enable such inductive testing. Data-driven research is not per se equivalent with inductive testing, which would require keen awareness of the translation that is required between real-world empirical phenomena (events, entities and relations) and training data. For instance, the curation of training data-sets requires streamlining to achieve interoperability between data from different sources, creating potential misalignment between empirical observations and their digital proxy. On top of that, the relevant data may not be available and replaced by proxies that exacerbate

[34] Artem Shmatko and others, 'Learning the Natural History of Human Disease with Generative Transformers' [2025] Nature 1 <https://www.nature.com/articles/s41586-025-09529-3> accessed 18 September 2025. Clive Cookson, 'New AI Model Predicts Susceptibility to over 1,000 Diseases' *Financial Times* (17 September 2025) <https://www.ft.com/content/598e07ec-954f-49b7-9bc5-ce77f9fff934> accessed 18 September 2025.

[35] Hildebrandt, 'Ground-Truthing in the European Health Data Space' (n 31).

[36] See for the most recent update: European Commission: Directorate-General for Communications Networks, Content and Technology, State of the digital decade 2025 – 2030 digital decade, European Commission, 2025, https://data.europa.eu/doi/10.2759/3316141

the problem of translating real-world phenomena into digital data. As Callon and Law demonstrated,[37] to quantify requires prior qualification, because the real-world entities, events or relations that are quantified must be qualified as the same to become calculable. This qualification requires empirical work.

In other work I have discussed the need for such empirical work in terms of Geertz's concept of explication 'as a conceptual tool focused on discernment and judgment rather than calculation and reckoning'.[38] Explication means digging deeper, unearthing complexities, uncovering ambiguities, hidden confounding factors, conceptual overlap and double causalities. Here, the concept of a theoretical framework, rather than a mere 'theory' or 'hypothesis', makes sense, because the patterns that can explain observable facts are not necessarily causal, as events or entities may be mutually constitutive or depend on a web of performative speech acts. For instance, the legal effect of a specific event or action is not a logical implication or a causal effect but the result of dedicated legally relevant speech acts that qualify something as, for instance murder, a legal subject, as creating a liability to pay compensation, etc.

This means, for instance, that legal technologies, including those involving generative AI, must be scrutinised with regard to their reliability, based on an inquiry into their implied theoretical assumptions and the construction of a falsifiable theory that can explain the functionality of these technologies (not necessarily their inner workings, which is another matter). In computer science, correlations are often seen in terms of either logic (as in logic programming)[39] or causality (as a subdomain of machine learning).[40] If, however, the relevant relations between facts, circumstances, legislation and case law is not causal or a matter of logical deduction but due to the performative nature of speech acts, the use of these technologies may present us with an unexpected or even hidden unreliability. If we could develop the hidden assumptions that inform the development of computational models, into a sufficiently detailed theoretical framework, then we may be able to tease out the implications of mistaking a performative speech act for a cause, or even a correlation. This could enable the falsification of the theoretical framework and pinpoint where and how these models 'get it wrong' with regard to real-world decision making.

[37] M Callon and Law J., 'On Qualculation, Agency, and Otherness' (2005) 23 Environment and Planning D: Society and Space 717.

[38] Quoted from the abstract of Mireille Hildebrandt, 'Qualification and Quantification in Machine Learning. From Explanation to Explication' (2022) 16 Sociologica 37 <https://sociologica.unibo.it/article/view/15845> accessed 12 September 2025.

[39] Robert Kowalski, 'Logic Programming' in Jörg H Siekmann (ed), *Handbook of the History of Logic*, vol 9 (North-Holland 2014) <https://www.sciencedirect.com/science/article/pii/B9780444516244500125> accessed 12 September 2025.

[40] Judea Pearl, *Causality: Models, Reasoning and Inference* (2nd edition, Cambridge University Press 2009).