Specifying Computational Compliance for AI: Blueprint for a New Research Domain

Bill Marino[*]
University of Cambridge
wlm27@cam.ac.uk

Nicholas D. Lane
University of Cambridge

Abstract

AI systems, we argue, will be unable to comply with AI regulation (AIR) at the necessary speed and scale using traditional, analogue methods of compliance. Rather, compliance with these regulations can only be achieved computationally, via algorithms that run across the life cycle of the AI systems, automatically steering them toward compliance in the face of dynamic conditions. Despite their (we would argue) inevitability, the research community has yet to specify exactly how these algorithms for computational AIR compliance should behave — or how we should measure their success. To fill this gap, we specify a set of design goals for such algorithms. In addition, we specify benchmarks for quantitatively measuring whether they satisfy these design goals. By delivering this blueprint, we hope to give shape to an important but uncrystallized new domain of research — and, in doing so, incite necessary investment in it.

1 Introduction

This paper rests on the provocative premise that the future of all legal compliance is computational.

As every aspect of our lives becomes digitized, even if our laws are still printed in dust-gathering tomes and stenciled on road signs, compliance with those laws will be wholly managed by the architectures of — and algorithms inside — the digital systems that suffuse our world.

The benefits of this computationally compliant future will be manifold. It will reduce the cost of compliance, removing a key barrier to market entry and thus fostering competition [Klapper et al., 2006]. It will permit “regulatory compliance in real time” [Bamidele, 2025], with violations mitigated as soon as they occur — and, often, before any harm is done. What is more, by removing the potential for human error, computational compliance will ensure *better* compliance, and a reality that hews closer to the letter of the laws that encode our societal values.

As Artificial Intelligence Regulation (AIR) takes shape worldwide [Alanoca et al., 2025], we arge that it can (and should) represent the turning point in this evolution. “Since AI is an algorithm,” argues one author, “then the method of its regulation should be the use of an algorithm comprising legal standards” [Szostek, 2021].

In this paper, we sketch a blueprint for fulfilling that vision. In particular, we specify exactly how such an algorithm — one that runs across the life cycle of an AI system, dynamically steering it towards AIR compliance in the face of variable conditions (e.g., data drift, post-deployment human feedback, changing laws, and more) — should behave. That is to say, we specify *design goals* for computational AI regulation compliance (CAIRC). What is more, we specify how we can quantitatively measure our progress towards achieving those design goals using benchmarks.

[*] Corresponding author.

The 3rd Workshop on Regulatable ML @NeurIPS2025.

Above all, our hope is that this work brings structure and a set of lucid North Stars for future investment in this nascent but increasingly crucial field of research.

2 Why Computational AIR Compliance Is Inevitable

In short, we believe the expansiveness and expense of AI regulation are on a collision course with the complexity, scale, and dynamicism of AI in the modern era. In this new reality, the manual, analog compliance solutions of the past will prove unsustainable and CAIRC will emerge as the only viable method of complying with AI regulation.

As mentioned, countries across the world are moving to regulate AI — often with very different outcomes [Sloane and Wüllhorst, 2025, Chun et al., 2024, Alanoca et al., 2025]. If the European Union's Artificial Intelligence Act (EU AI Act) [European Union, 2024] (dubbed "the world's first comprehensive AI law" [European Parliament, 2024]) is any indication, then these regulations will have an "expansive scope" [Addey, 2023]: reaching deep into the details of AI systems and models (collectively, "AI") to dictate "complex rules"[Zulehner, 2024] around everything from their training data to their performance levels, logging practices, and more [European Union, 2024, Art. 10, 12, 15]. If the EU AI Act is any indication, complying with these regulations will also carry considerable expense for the regulated [Wu and Liu, 2023] — perhaps even cost-prohibitive expense in the case of small and -medium size enterprises [Schneier and Sanders, 2023, Gikay, 2024, Wu and Liu, 2023, Government, 2023, Haataja and Bryson, 2021, Sullivan, 2024, Reuel et al., 2024b, Koh et al., 2024, Bolda, 2024, Molnar, 2024][2].

Meanwhile, on the other side of the equation is a "brave new world of AI" [Vithayathil and Nauroth, 2023] that is more complex, dynamic, scaled-up, and global than ever before. The complexity of today's AI [Zaharia et al., 2024] — as well as the development pipeline [Sadek et al., 2024] and supply chain behind it [Brown, 2023, Engler and Renda, 2022, Marino et al., 2024] — is at an all-time high. AI systems and models today often comprise dozens of datasets and models, many externally sourced from third parties via API or community platforms like Hugging Face. [Amershi et al., 2019, Take et al., 2021, Chaudhuri et al., 2024, Renieris et al., 2023, Osborne et al., 2024, Jones et al., 2024, Ada Lovelace Institute, 2023, Liesenfeld and Dingemanse, 2024, Barclay et al., 2019]. Meanwhile, the training datasets for some models are nearing "unimaginable scale" [Coders Stop, 2025, Shen et al., 2025]; by 2028, training sets are expected to "approach[] the total effective stock of text in the indexed web" [Villalobos et al., 2024]. As we consider a near future where AI systems include "hundreds of agents" [Falconer, 2025], this complexity may only increase. Adding fuel to the fire is the fact that "AI systems are constantly changing and evolving" [Nicenboim et al., 2022]. Specifically, they are the product of "continuous experimentation"[Martínez-Fernández et al., 2022] and "agile" software development processes that prize "rapid iteration[]" in response to changing "customer needs, technical changes, and market volatility" [Balayn and Gürses, 2024, Carlini, 2022, Xin et al., 2018, Guo et al., 2024, Piorkowski et al., 2022] — as well as "continual learning" methods [Wang et al., 2024] whereby production data is continually used to retrain and improve the AI. Last but not least, AI is increasingly marketed toward an international audience [Organization, 2024, Reuters, 2025], in which case they must comply with the entire patchwork of AI regulations described before.

The net takeaway is that AI — either today or, at least, in the near future — may simply be too complex, dynamic, large, and global for the traditional, human-driven models of regulatory compliance [O'Reilly, 2025, Krasadakis, 2023, Marino et al., 2024, Marino, 2024, Anderljung et al., 2023, Hacker et al., 2023, Confino, 2024, Fiazza, 2021]. That is, human compliance practitioners will be unable to handle the task of determining whether complicated and ever-changing AI of titanic scale comply with a protean patchwork of AIR — or, if they do not, determining how to bring them back into a compliant state. This will leave no choice but to shift to AIR compliance methods that are as scalable and dynamic as their AI subjects — i.e., computational.

3 Deconstructing the problem

> "If you're overwhelmed by the whole, break it down into pieces" — Chuck Close [Ward, 2007]

[2] EU AI Act compliance costs for some types of AI systems, for example, are estimated to be as high as €400,000 [Koh et al., 2024, 1872]

When developing algorithms for CAIRC, what should our design goals be? And how do we quantitatively measure our progress toward them?

To help answer these questions, we find it useful to deconstruct CAIRC into two sub-problems. Specifically, we posit that any CAIRC algorithm must necessarily contain two complimentary functions, which we deem the *Inspector* and the *Mechanic*:[3]

As depicted in Fig. 1, the *Inspector* will diagnose — at any given point in time and in a fully automated manner — the AIR compliance level of an AI. When it finds that the AI is not compliant with one or more AIRs, it will communicate its diagnosis to the *Mechanic*, which will endeavor to remedy the non-compliance using various automated tools, ultimately calling on the *Inspector* to re-run its audit and determine if a compliance state has been achieved (or, perhaps, restored).

In the sections that follow, we set design goals and benchmarking criteria for each of these two functions — as well as the broader CAIRC algorithm that necessarily unites and envelopes them.

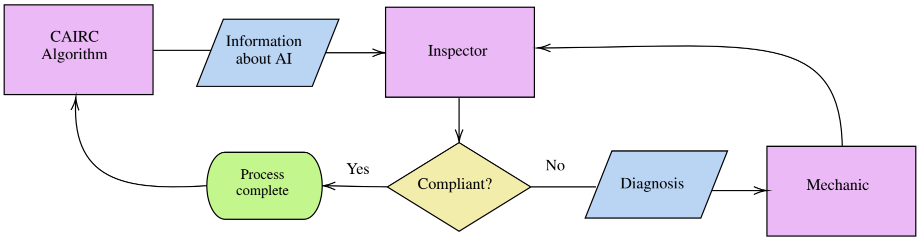

Figure 1: **CAIRC flowchart**. As a first step, the overarching CAIRC algorithm submits information about an AI to the *Inspector* (e.g., as a scheduled job). Next, the *Inspector* reaches a finding of either compliance, in which case the process is complete, or non-compliance, in which case the *Inspector* transmits its diagnosis to the *Mechanic*. Upon receiving it, the *Mechanic* uses its tools to try to repair the diagnosed compliance defect(s). When finished, it calls the *Inspector* to re-run its analysis. This loop repeats until the *Inspector* finds that compliance exists, in which case the process is completed (a fact that is communicated back to the overarching CAIRC algorithm, so that it can adjust its schedule accordingly).

4 The *Inspector*

In this section, we lay out the design criteria that, we argue, a CAIRC algorithm's *Inspector* function must satisfy. We also describe methods for benchmarking an *Inspector*, to quantitatively measure whether those design criteria are satisfied.

4.1 Design Criteria

Our position is that an *Inspector*, in order to fulfill its purpose, must satisfy several key design criteria. These relate to:

- The *Inspector*'s input;
- The *Inspector*'s output;
- The function that maps the former to the latter.

Below, we describe these design criteria in detail. Where applicable, we describe how close the current state of the art (SOTA) comes to achieving these design goals and/or identify any open research problems that must be solved before these design criteria can realistically be satisfied.

[3] Happily, the *Inspector* and *Mechanic* have independent, standalone value. Even in the absence of a *Mechanic* to automatically repair the compliance defects it identifies, the *Inspector* can be used to alert human compliance assessors or "human *Mechanics*" to compliance defects. Conversely, the *Mechanic* can be used to cure defects identified by humans.

4.1.1 Input

In order to assess the AIR compliance level of a given AI, the *Inspector* requires, as its input, information about that AI. Importantly, this information — and therefore the *Inspector* input — must satisfy the following design criteria:

Comprehensiveness : If an *Inspector* is to accurately and holistically assess the AIR compliance of an AI, then the information inputted into it must describe *all* aspects of the AI that bear (or could potentially bear) on that compliance. Failure to input all of the information relevant to AIR compliance carries great risk: specifically, of false positives (FP), whereby the *Inspector* incorrectly labels a non-compliant AI compliant because it is not privy to the factual information indicating otherwise. Because FPs like these could lead to penalties [European Union, 2024, Art. 99] and even harm (of the sort the AIR aims to prevent), they must be avoided. And the only way to do that is to ensure the *Inspector* inputs cover *all* aspects of the AI that bear on its AIR compliance.

So, for example, information relevant to EU AI Act compliance might concern everything from an AI system's data governance practices [European Union, 2024, Art. 10] and human oversight mechanisms [European Union, 2024, Art. 14], which are the direct subjects of EU AI Act requirements. But it will also necessarily include information about that AI system's intended use, which determines the particular set of rules that apply to it [European Union, 2024, Art. 6], and whether it is open source, which potentially exempts it from those rules [European Union, 2024, Art. 2]. The input to the *Inspector* must therefore include the super set of all this information — and any other information relevant to EU AI Act compliance.

Importantly, this comprehensiveness must be achieved *for every AIR that the system is expected to comply with*. Given the increasingly global nature of AI, this may mean dozens of AIR for a given AI system. In these cases, the super set of information relevant to each and every AIR must be inputted into the *Inspector*.

Attestability : Information that is relevant to an AI system's AIR compliance may go beyond information about that particular system or model, to its ingredient models, datasets, and more. The EU AI Act, for example, includes a number of requirements around training data [European Union, 2024, Art. 10]. In today's complex AI supply chain, this training data may come from disparate sources, including non-trusted providers via API or online communities like Hugging Face [Marino et al., 2024]. In these cases, it will be crucial to verify, sometimes without direct access to the subject of the verification (i.e., through "remote attestation" [Brundage et al., 2020]) that the information about the training data that is inputted into the *Inspector* is accurate [Marino, 2024, Reuel et al., 2024a]. At the moment, this type of attestation is considered an "open problem" [Reuel et al., 2024a], but various methods are being explored [Cen and Alur, 2024, South et al., 2024, Sun and Zhang, 2023, Hugging Face, 2024, Schnabl et al., 2025].

Concurrency : To achieve true CAIRC, the input must reflect the current state of the AI system. In other words, the *Inspector* must have up-to-date knowledge of all AIR-relevant facets of the system, including dynamic facets like logs, user feedback, cybersecurity attacks, and more. Information that is outdated — even by seconds — represents a grave FP risk.

4.1.2 Output

When the *Inspector* finds that AIR compliance exists, it need not output anything other than, perhaps, a void return. In all other cases, the key design criteria for the *Inspector* output is that it provide enough information for the *Mechanic* to fulfill its role of repairing any identified compliance deficiencies and achieving or restoring compliance to the AI (i.e., is "*Mechanic*-enabling").

Among other things, this means that the *Inspector*'s cannot simply return a binary class label of "non-compliant" or, differently, a single aggregate compliance score [Guldimann et al., 2024]. At a minimum, what is required are outputs that are granular (high fidelity) enough that the *Mechanic* knows what work to begin *and where* — without, in the interests of efficiency, needing to duplicate any of the compliance assessment work done by the *Inspector*. For example, in communicating a violation of Article 10 of the EU AI Act, the *Inspector* would probably need to include, in its output, a dataset identifier along with the particular section of Article 10 that was violated.

Where an *Inspector* with deeper access to a system (e.g., individual data points in a training set) has surfaced more granular compliance violation information in performing its assessment, it may transmit this additional information (e.g., data point identifiers) to the *Mechanic*, to relieve it of the task of pinpointing the exact sources of non-compliance.[4]

4.1.3 Function Mapping Input to Output

The final cornerstone of the *Inspector* is some function that accurately maps its input onto its output; i.e., maps information about an AI onto a *Mechanic*-enabling AIR compliance diagnosis. The function could consist of an LLM [Sovrano et al., 2025, Li et al., 2025, Makovec et al., 2024], rule-based algorithm [Marino et al., 2024], evaluation suites that run on AI assets [Sovrano and Vitali, 2023, Walke et al., 2023, Nolte et al., 2024, Bueno Momcilovic et al., 2024, Esiobu et al., 2023, Qin et al., 2023, Lin et al., 2022, Parrish et al., 2022, Guldimann et al., 2024, Chen et al., 2024], combinations of these, or anything else.

Regardless of this mapping function's exact contents, it must accurately map inputs onto outputs; i.e., map information about AI systems and models onto accurate compliance predictions. Because FPs (findings of compliance when an AI is, in fact, non-compliant) are especially costly in this setting, it must have a low FP rate; i.e., high precision.

4.2 Benchmark

To quantitatively measure our progress toward these design goals, we need to be able to benchmark the ability of proposed *Inspector* algorithms to successfully predict the compliance level of a given AI system at a given point in time, in light of one or more AIR. A benchmark dataset that would fill this gap might consist of whole *Inspector* inputs — i.e., sets of information about AI systems, satisfying our input design criteria above — labeled by ground truth outputs — i.e., compliance diagnoses. Such a benchmark could be used to evaluate the accuracy with which candidate *Inspector* algorithms predict the ground truth, as well as the speed and cost at which they do it (if it compares to the speed of manual compliance analyses, then this undermines some of the benefits of CAIRC put forth in Sec. 1).[5] Notably, despite the growing number of algorithms in the literature that, like our proposed *Inspector*, automatically assess the AIR compliance of an AI (cataloged in Sec. 4.1.3), only one benchmark dataset for measuring the performance of these algorithms currently exists — and it is strictly focused on LLM-based approaches [Marino et al., 2025a].

5 The *Mechanic*

In this section, we lay out design criteria for the *Mechanic* function. We also describe a method for benchmarking the *Mechanic*, to quantitatively measure whether those design criteria are being achieved.

5.1 Design Criteria

Our position is that a *Mechanic*, in order to fulfill its function, must satisfy several key design criteria. These relate to:

- The *Mechanic*'s input;
- The *Mechanics*'s output;
- The repair algorithm(s) employed by the *Mechanic*.

Below, we describe these in more detail. Where applicable, we refer to the SOTA as well as any open research problems that must be solved before these design criteria can realistically be satisfied.

[4] Note that there may often be reason to keep some aspects of the AI out of the hands of the *Inspector* — for example, if the *Inspector* is being operated by an arms-length auditor or a regulator (an arrangement would could have benefits in terms of providing an external check on the AI). In these situations, the *Inspector* may not, by design, have access to enough information about the AI to provide a granular output to the *Mechanic*.

[5] The challenge of creating the ground truth for such a benchmark should not be underestimated. Compliance, it has been said, is "hard to measure" and "not binary" [Wu and van Rooij, 2021]. In creating ground truth, it will be important to account for "grey areas."

5.1.1 Input

The *Mechanic* must accept, as its input, the output of the *Inspector* (whose design criteria were described in Sec. 4.1.1). As previously discussed, the granularity of this input may influence the scope of the *Mechanic*'s functionality and the details of its internal algorithm (covered in 5.1.3).

5.1.2 Output

The *Mechanic* (or, more specifically, its repair algorithm described below) are tasked with making repairs directly to the AI. This includes making changes to the AI's assets: its code, data, models, documentation, and more. On one hand, the output of the *Mechanic* is the altered version of these assets (e.g., the data it has filered, the models it has re-trained, etc.). More concretely, the *Mechanic* should also output a signal (e.g., a void function return) that indicates that its work, from its point of view, is complete. Upon receiving this signal, the overarching algorithm that encompasses the *Mechanic* and the *Inspector* can call on the *Inspector* again, to check the *Mechanic*'s work (i.e., to verify whether compliance has in fact been achieved).

5.1.3 Repair algorithm

What lies between the input and the output of the *Mechanic* is a repair algorithm or program that must accomplish a few key tasks:

Pinpoint the non-compliance (optional) : Depending on the particular AIR violation as well as the granularity of the *Inspector* output, the *Mechanic* may need to do additional legwork to pinpoint the exact source of the non-compliance (e.g., identify the data points deemed to be causing unmitigated data poisoning in violation of European Union [2024, Art. 15]). Put differently, where the outputs of the *Inspector* are sparse, the *Mechanic* must possess the functionality to discretely scan the AI for the sources of non-compliance — or to otherwise map high-level compliance violation descriptions onto the atomic components of the system that must be repaired.

Select the tool(s) to repair the non-compliance We define the *Mechanic*'s tools as those discrete functions that the *Mechanic*'s repair algorithm will call upon in order to execute repairs to the AI's sources of non-compliance and bring the AI back to a compliant state.[6] Here, for example, is a non-complete list of sample tools a *Mechanic* might want to have at its disposal in order to repair various AIR deficiencies:

- Where non-compliance stems from biased (and unmitigated) outputs of a generative AI model [European Union, 2024, Art. 9, 55], the *Mechanic* may leverage a machine unlearning tool [Cao and Yang, 2015, Hine et al., 2024, Xu et al., 2024, Marino et al., 2025b], a model editing [Gupta et al., 2024] tool, or a [Qi et al., 2023] tool, to try to suppress the biased outputs without the need for full retraining of the model.
- Where non-compliance stems from model inaccuracy [European Union, 2024, Art. 15], the *Mechanic* may leverage tools for improving accuracy by acquiring (and then re-training on) more or better data from new sources; this, in turn, may require the ability to generate synthetic data [Bauer et al., 2024] or buy it on data marketplaces, to label, filter, or otherwise prepare that data for training, and, lastly, to retrain and evaluate the downstream model.
- Where non-compliance stems from model leakage of personal data in the training set [European Union, 2024, Art. 14], the *Mechanic* may require access to a differential privacy (DP) tool [Bauer et al., 2024, Marino et al., 2025b]) that it can apply before retraining in order to mitigate the risk of leakage in the model;

These tools must have the ability to edit the AI system: e.g., filter training sets, retrain models, and more. The *Mechanic*, meanwhile, must possess the ability to map *Inspector* outputs onto the right tools (e.g., through rule-based methods or by relying on an LLM to reason about which tools to

[6] Tools is a popular term in the world of AI agents, where it refers to those utilities that help connect an LLM to external resources like internet browsers [Wiesinger et al., 2025, Ruan et al., 2023, Woodside and Toner, 2024], and it re-use here is not purely coincidental. This is because is not hard to imagine an agentic implementation of CAIRC where the *Inspector* and *Mechanic* are subagents and the *Mechanic*'s tools are agentic tools (or perhaps other subagents).

leverage [Microsoft, 2024]) and also to navigate trade-offs between different tool options based on things like cost, latency, and ability to cure the particular defect at hand.

There is work to be done mapping out the full spectrum of tools required by the *Mechanic* to bring the AI system, under any scenario, back to a compliant state. Importantly, to achieve true CAIRC, the *Mechanic* algorithm must have access to a set of tools that, working together, can solve any arbitrary AIR compliance deficiency. At the outset, we should highlight the fact that we do not believe this full set of tools exist yet in the SOTA. In particular, we can assume that no tools yet exist wherever, in the eyes of scholars, AIR calls for "technical capabilities or engineering solutions that do not currently exist" [Guha et al.] or otherwise "rest on open issues in computer science" [Fiazza, 2021], including around transparency [Guha et al.], human oversight [Ebers et al., 2021], data quality [Ebers et al., 2021, Heikkilä, 2022, Microsoft, 2021, Fiazza, 2021, Microsoft, 2021, e Silva, 2024], and the robustness, explainability, and security of models [Fiazza, 2021, Guha et al., Heikkilä, 2022, Marino, 2024, Morley et al., 2020, Marino, 2024].

Orchestrate and manage the execution of those tools, through to some predicted state of completion Once it has selected the specific tool(s) that it will use to address the non-compliance, the *Mechanic* repair algorithm must orchestrate and manage the use of those tools to cure the particular deficiency. This includes the ability to monitor the progress and efficacy of these orchestrated tools – i.e., as well as make a preliminary prediction about whether the tool has resolved the non-compliance (and, therefore, whether it is time to send an output message to the overarching algorithm that encompasses the *Mechanic* and the *Inspector*).

5.2 Benchmark

To quantitatively measure our progress toward these design goals, we need to be able to benchmark the ability of proposed *Mechanic* algorithms to effectively repair AIR compliance defects in an AI. A benchmark dataset would help. Such a benchmark dataset might consist of AI systems or models that are non-compliant with one or more AIR, ideally in different ways. The full suite of assets comprising each AI would be included in the dataset: that is to say, their complete training and evaluation datasets, their model weights, and their training, evaluation, and deployment code (i.e., full "snapshots"). In addition, each AI would be labeled with, essentially, an *Inspector* output (or other report card) that includes a diagnosis of the particular compliance issue. *Mechanic* algorithms should be fed the label and asked to operate on the AIs assets in order to repair the diagnosed compliance defect. Optionally, it could also be given the ability to call the an *Inspector* to evaluate its repairs. *Mechanic* algorithms could be evaluated for their success rate in being able to achieve a compliant state, as graded by the *Inspector* — as well as the number of calls to *Inspector* required to get there and the speed or computational cost in doing so.[7]

6 Connecting the *Inspector* and *Mechanic* in a Closed-loop System

The *Inspector* and *Mechanic* should ultimately be connected and encompassed by an overarching algorithm, creating a single, unified system for CAIRC. This closed-loop system will need to manage the following:

1. Run the *Inspector* routinely, perhaps as a scheduled job and ideally with enough frequency that AIR violations are detected and eliminated before harm is caused;
2. Route non-void *Inspector* outputs (i.e., findings of non-compliance) to the *Mechanic*;
3. When the *Mechanic* returns, re-run the *Inspector*;
4. Repeat this loop until the *Inspector* returns void (indicating compliance has been restored);

It is important to note that this unified system could, in theory, be split across multiple organizations. For example, the *Mechanic* could be owned by an AI developer while the *Inspector* could belong to an auditing company or even regulator. This would permit an external check on the compliance levels of the AI — without given external entities access to certain parts of the AI system.

[7] Note that measuring speed and cost is important because it not only helps us compare *Mechanic* algorithms, but helps us compare *Mechanic* algorithms with human-driven compliance protocols. This might, in turn, support the hypothesis, put forth in Sec. 1, that CAIRC can lower costs compared to human-driven compliance efforts.

The overarching algorithm must also have the ability to detect an endless loop between the *Mechanic* and the *Inspector*, possibly triggering more severe mitigations, such as a pause of the AI system.

6.1 Benchmark

Although benchmarking the *Inspector* and *Mechanic* algorithms independently is valuable, it will also be important to benchmark the close-loop CAIRC system that envelopes them. This will help us test the way they behave together, including how often they enter an endless loop and, working together, fail to cure a given AIR compliance deficiency. A benchmark dataset for testing the complete CAIRC system might consist, like the *Inspector* benchmark, of whole *Inspector* inputs — i.e., sets of information about AI systems, satisfying our input design criteria above — labeled by ground truth outputs — i.e., compliance diagnoses. After the CAIRC has run its course, and the *Mechanic* has made its changes to the AI, human experts could qualitatively check whether the resulting AI is indeed compliant. Or, differently, a SOTA LLM that has already proven to be effective at the *Inspector* task could be used, as a model-as-judge [Gu et al., 2025], to assess the AIR compliance level of the resulting AI. Separately, the rate of failures (where the *Inspector* and *Mechanic* get caught in an endless loop), as well as the speed and cost of the end-to-end system, could be tracked.

7 Challenges

Computationality aside, AIR compliance is haunted by existential questions about its technical feasibility and measurability [Guha et al., 2024, Guha et al.]. Critics argue that compliance with the EU AI Act, for example, rests on a number of open problems around explainability, human oversight, cybersecurity, and more [Guha et al., 2024, Fiazza, 2021, Guha et al., Ebers et al., 2021, Heikkilä, 2022, Microsoft, 2021, Fiazza, 2021, Microsoft, 2021, e Silva, 2024, Heikkilä, 2022, Marino, 2024, Morley et al., 2020, Marino, 2024]. Differently, it has been said that EU AI Act compliance will be difficult or even impossible to measure [Almada and Petit, 2023] due to a lack of agreed-upon benchmarks for core concepts like bias [Committee on Standards in Public Life, 2020, Buyl and Bie, 2024, Dulka, 2023, Gornet, 2024] and interpretability [Guha et al., Hutson, 2023]. With LLMs in particular it has been said that it is "impossible to demonstrate compliance with a given regulatory specification" [Judge et al., 2024, Saeed and Omlin, 2023, Lee et al., 2024]. These critiques foreshadow potential hurdles en route to CAIRC, of course, because if researchers have not yet figure out how to measure or execute compliance in certain AIR scenarios, how can we expect our *Inspector* and *Mechanic* to do so?

As a separate matter, when it comes to compliance, there are those that hold the viewpoint that "[h]uman oversight, nuanced judgment, ethical considerations, and strategic thinking cannot, and should not, be outsourced entirely to algorithms" [Compliance Podcast Network, 2025]. This may stem from the notion that compliance, general, is "hard to measure" and "not binary" [Wu and van Rooij, 2021]. Needless to say, making AIR compliance computational (and especially benchmarking it) requires the opposite view: that compliance can successfully be encoded in digital systems that must make, in some cases, binary predictions — with their performance quantitatively measured using objective ground truth. If and when "grey areas" emerge in the application of AIR, this threatens the value and viability of CAIRC. Accordingly, it is a risk worth monitoring closely as we develop CAIRC algorithms.

8 Conclusion

Legal compliance, we argue, will ultimately be governed not by human oversight but by algorithms operating within digital systems — making it inherently computational. AI regulation represents a prime opportunity to begin that transition. To move the field forward, we propose a set of design principles to steer the development of computational AIR compliance algorithms and, additionally, introduce benchmarks to quantitatively measure how faithfully those algorithms meet the design principles. Our intention in laying out this framework is to help crystallize a research area that is still being formed, while also sparking additional research investment in it.

References

Ada Lovelace Institute. What is a foundation model? https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/, 2023. [Accessed 22-10-2025].

Mark Addey. Charting a new era: the European Union's AI legislation and its transformative influence on technology and society. *SSRN Electronic Journal*, 2023. doi: 10.2139/ssrn.4560262. URL https://ssrn.com/abstract=4560262.

Sacha Alanoca, Shira Gur-Arieh, Tom Zick, and Kevin Klyman. Comparing apples to oranges: A taxonomy for navigating the global landscape of AI regulation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 914–937. ACM, June 2025. doi: 10.1145/3715275.3732059. URL http://dx.doi.org/10.1145/3715275.3732059.

Marco Almada and Nicolas Petit. The EU AI act: A medley of product safety and fundamental rights? Working Paper 2023/59, European University Institute, 2023. URL https://hdl.handle.net/1814/75982.

Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In Helen Sharp and Mike Whalen, editors, *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE (SEIP) 2019, Montreal, QC, Canada, May 25-31, 2019*, pages 291–300. IEEE / ACM, 2019. doi: 10.1109/ICSE-SEIP.2019.00042. URL https://doi.org/10.1109/ICSE-SEIP.2019.00042.

Markus Anderljung, Emma Barnhart, Anton Korinek, Jeffrey Leung, Cullen O'Keefe, Jess Whittlestone, et al. Frontier AI regulation: Managing emerging risks to public safety. Unpublished manuscript, 2023.

Agathe Balayn and Seda Gürses. Misguided: AI regulation needs a shift in focus. *Internet Policy Review*, 13(3), September 2024. URL https://policyreview.info/articles/news/misguided-ai-regulation-needs-shift/1796. Open access opinion piece.

Matthew Bamidele. Integration of AI with IoT for real-time compliance in connected insurance. *ResearchGate*, August 2025. URL https://www.researchgate.net/publication/394753646_Integration_of_AI_with_IoT_for_Real-Time_Compliance_in_Connected_Insurance. Uploaded 20 August 2025.

Iain Barclay, Alun D. Preece, Ian J. Taylor, and Dinesh C. Verma. Quantifying transparency of machine learning systems through analysis of contributions. *CoRR*, abs/1907.03483, 2019. URL http://arxiv.org/abs/1907.03483.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey, 2024. URL https://arxiv.org/abs/2401.02524.

Stefan Bolda. Navigating the EU AI act: Proposed compliance measures for AI providers and deployers. Master's thesis, Johannes Kepler University Linz, Linz, Austria, 2024. URL urn:nbn:at:at-ubl:1-80988. Thesis advisor: Barbara Krumay.

Ian Brown. Allocating accountability in AI supply chains. https://www.adalovelaceinstitute.org/resource/ai-supply-chains/, 2023. [Accessed 22-10-2025].

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020. URL https://arxiv.org/abs/2004.07213.

Tomas Bueno Momcilovic, Beat Buesser, Giulio Zizzo, Mark Purcell, and Dian Balta. Assuring compliance of LLMs with EU AIA robustness demands. In *Wirtschaftsinformatik 2024 Proceedings*, page 126, 2024. URL https://aisel.aisnet.org/wi2024/126.

Maarten Buyl and Tijl De Bie. Inherent limitations of AI fairness. *Commun. ACM*, 67(2):48–55, 2024. doi: 10.1145/3624700. URL https://doi.org/10.1145/3624700.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. doi: 10.1109/SP.2015.35.

Nicholas Carlini. Rapid iteration in machine learning research. https://nicholas.carlini.com/writing/2022/rapid-iteration-machine-learning-research.html, 2022. [Accessed 22-10-2025].

Sarah H. Cen and Rohan Alur. From transparency to accountability and back: A discussion of access and evidence in AI auditing, 2024. URL https://arxiv.org/abs/2410.04772.

Shamik Chaudhuri, Kingshuk Dasgupta, Michael Le Isaac Hepworth, Mark Lodato, Mihai Maruseac, Sarah Meiklejohn, Tehila Minkus, and Kara Olive. Securing the AI software supply chain. https://research.google/pubs/securing-the-ai-software-supply-chain/, 2024. [Accessed 22-08-2025].

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation, 2024. URL https://arxiv.org/abs/2407.07087.

Jon Chun, Christian Schroeder de Witt, and Katherine Elkins. Comparative global AI regulation: Policy perspectives from the EU, China, and the US, 2024. URL https://arxiv.org/abs/2410.21279.

Coders Stop. The inconvenient truth about AI training data that companies are hiding, July 2025. URL https://medium.com/@coders.stop/the-inconvenient-truth-about-ai-training-data-that-companies-are-hiding-1a3545993164.

Committee on Standards in Public Life. Artificial intelligence and public standards: A review by the Committee on Standards in Public Life. Government review, Government of the United Kingdom, February 2020. URL https://assets.publishing.service.gov.uk/media/5e553b3486650c10ec300a0c/Web_Version_AI_and_Public_Standards.PDF. Chair: Lord Evans of Weardale KCB DL.

Compliance Podcast Network. Stepping up and stepping forward: The future of compliance in an age of AI and deregulation, April 2025. URL https://compliancepodcastnetwork.net/stepping-up-and-stepping-forward-the-future-of-compliance-in-an-age-of-ai-and-deregulation/. [Accessed 22-10-2025].

Paolo Confino. Tom siebel: Ai models are too complex for regulators—new government agencies won't help. *Yahoo Finance*, September 2024. URL https://finance.yahoo.com/news/tom-siebel-ai-models-too-091000461.html. Interview on regulatory challenges concerning AI model complexity.

Anne Dulka. The use of artificial intelligence in international human rights law. *Stanford Technology Law Review*, 26:316, 2023.

Nuno Sousa e Silva. The Artificial Intelligence Act: Critical overview. *CoRR*, abs/2409.00264, 2024. doi: 10.48550/ARXIV.2409.00264. URL https://doi.org/10.48550/arXiv.2409.00264.

Martin Ebers, Veronica R. S. Hoch, Frank Rosenkranz, Hannah Ruschemeier, and Björn Steinrötter. The European Commission's proposal for an Artificial Intelligence Act—a critical assessment by members of the Robotics and AI Law Society (RAILS). *J*, 4(4):589–603, 2021. ISSN 2571-8800. doi: 10.3390/j4040043. URL https://www.mdpi.com/2571-8800/4/4/43.

Alex Engler and Andrea Renda. Reconciling the AI value chain with the EU's Artificial Intelligence Act. https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/, 2022. [Accessed 22-10-2025].

David Esiobu, Xiaoqing Ellen Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. ROBBIE: Robust bias evaluation of large generative language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3764–3814. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.230. URL https://doi.org/10.18653/v1/2023.emnlp-main.230.

European Parliament. EU AI Act: First regulation on artificial intelligence. *European Parliament Topics*, June 2024. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

European Union. Artificial Intelligence Act, March 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Official Journal of the European Union.

Sean Falconer. More than machines: The inner workings of AI agents, March 2025. URL https://seanfalconer.medium.com/more-than-machines-the-inner-workings-of-ai-agents-5bba7904d04e.

Maria-Camilla Fiazza. The EU proposal for regulating AI: Foreseeable impact on medical robotics. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 222–227, 2021. doi: 10.1109/ICAR53236.2021.9659429.

Asress Adimi Gikay. Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation. *International Journal of Law and Information Technology*, 32(1):eaae013, 06 2024. ISSN 0967-0769. doi: 10.1093/ijlit/eaae013. URL https://doi.org/10.1093/ijlit/eaae013.

Mélanie Gornet. The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation. Technical report, 2024. ffhal-04785519f.

UK Government. UK Artificial Intelligence Regulation Impact Assessment. https://assets.publishing.service.gov.uk/media/6424208f3d885d000cdadddf/uk_ai_regulation_impact_assessment.pdf, 2023. [Accessed 22-08-2025].

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, and Daniel E. Ho. The AI regulatory alignment problem. https://hai.stanford.edu/sites/default/files/2023-11/AI-Regulatory-Alignment.pdf. [Accessed 22-08-2025].

Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, and Daniel E. Ho. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review*, 92(6):1473, 2024.

Philipp Guldimann, Alexander Spiridonov, Robin Staab, Nikola Jovanović, Mark Vero, Velko Vechev, Anna Gueorguieva, Mislav Balunović, Nikola Konstantinov, Pavol Bielik, Petar Tsankov, and Martin Vechev. COMPL-AI framework: A technical interpretation and LLM benchmarking suite for the EU Artificial Intelligence Act, 2024. URL https://arxiv.org/abs/2410.07959.

Grace Guo, Dustin Arendt, and Alex Endert. Explainability in JupyterLab and beyond: Interactive XAI systems for integrated and collaborative workflows. *CoRR*, abs/2404.02081, 2024. doi: 10.48550/ARXIV.2404.02081. URL https://doi.org/10.48550/arXiv.2404.02081.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. A unified framework for model editing, 2024. URL https://arxiv.org/abs/2403.14236.

Meeri Haataja and Joanna J. Bryson. What costs should we expect from the EU's AI Act? SocArXiv 8nzb4, Center for Open Science, August 2021. URL https://ideas.repec.org/p/osf/socarx/8nzb4.html.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating ChatGPT and other large generative AI models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 14, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3593013.3594067.

Melissa Heikkilä. A quick guide to the most important AI law you've never heard of. https://www.technologyreview.com/2022/05/13/1052223/guide-ai-act-europe/, 2022. [Accessed 22-08-2025].

Emmie Hine, Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Supporting trustworthy AI through machine unlearning. *Sci. Eng. Ethics*, 30(5):43, 2024. doi: 10.1007/S11948-024-00500-5. URL https://doi.org/10.1007/s11948-024-00500-5.

Hugging Face. Add verifyToken field to verify evaluation results are produced by Hugging Face's automatic model evaluator. https://huggingface.co/facebook/bart-large-cnn/discussions/23, 2024. [Accessed 22-10-2025].

Matthew Hutson. Rules to keep AI in check: nations carve different paths for tech regulation. *Nature*, 620 (7973):260–263, August 2023. doi: 10.1038/d41586-023-02491-y. PMID: 37553464.

Jason Jones, Wenxin Jiang, Nicholas Synovic, George K. Thiruvathukal, and James C. Davis. What do we know about Hugging Face? A systematic literature review and quantitative validation of qualitative claims. *CoRR*, abs/2406.08205, 2024. doi: 10.48550/ARXIV.2406.08205. URL https://doi.org/10.48550/arXiv.2406.08205.

Brian Judge, Mark Nitzberg, and Stuart Russell. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, page puae020, 05 2024. ISSN 1449-4035. doi: 10.1093/polsoc/puae020. URL https://doi.org/10.1093/polsoc/puae020.

Leora Klapper, Luc Laeven, and Raghuram Rajan. Entry regulation as a barrier to entrepreneurship. *Journal of Financial Economics*, 82(3):591–629, 2006. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2005.09.006. URL https://www.sciencedirect.com/science/article/pii/S0304405X06000936.

Florence Koh, Kathrin Grosse, and Giovanni Apruzzese. Voices from the frontline: Revealing the AI practitioners' viewpoint on the European AI Act. In *Proceedings of the Hawaii International Conference on System Sciences*, HICSS, 2024.

George Krasadakis. To regulate or not? How should governments react to the AI revolution?, October 2023. URL https://medium.com/60-leaders/to-regulate-or-not-how-should-governments-react-to-the-ai-revolution-c254d176304f. 32 min read.

Donghyeok Lee, Christina Todorova, and Alireza Dehghani. Ethical risks and future direction in building trust for large language models application under the EU AI Act. pages 41–46, 12 2024. doi: 10.1145/3701268.3701272.

Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. PrivaCI-Bench: Evaluating privacy with contextual integrity and legal compliance, 2025. URL https://arxiv.org/abs/2502.17041.

Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative AI: open washing and the EU AI Act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 1774–1787. ACM, 2024. doi: 10.1145/3630106.3659005. URL https://doi.org/10.1145/3630106.3659005.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Barbara Makovec, Luis Rei, and Inna Novalija. Preparing AI for compliance: Initial steps of a framework for teaching LLMs to reason about compliance. In *Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning (RuleML+RR'24)*, volume 3816, Bucharest, Romania, September 2024. CEUR Workshop Proceedings. URL https://ceur-ws.org/Vol-3816/paper63.pdf.

Bill Marino. The EU AI Act's technical "tension areas". https://www.lcfi.ac.uk/news-events/blog/post/the-eu-ai-acts-technical-tension-areas, 2024. [Accessed 22-10-2025].

Bill Marino, Yaqub Chaudhary, Yulu Pi, Rui-Jie Yew, Preslav Aleksandrov, Carwyn Rahman, William F. Shen, Isaac Robinson, and Nicholas D. Lane. Compliance Cards: Automated EU AI Act compliance analyses amidst a complex AI supply chain, 2024. URL https://arxiv.org/abs/2406.14758.

Bill Marino, Rosco Hunter, Zubair Jamali, Marinos Emmanouil Kalpakos, Mudra Kashyap, Isaiah Hinton, Alexa Hanson, Maahum Nazir, Christoph Schnabl, Felix Steffek, Hongkai Wen, and Nicholas D. Lane. AIReg-Bench: Benchmarking language models that assess AI regulation compliance, 2025a. URL https://arxiv.org/abs/2510.01474.

Bill Marino, Meghdad Kurmanji, and Nicholas D. Lane. Bridge the gaps between machine unlearning and AI regulation, 2025b. URL https://arxiv.org/abs/2502.12430.

Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, and Stefan Wagner. Software engineering for AI-based systems: A survey. *ACM Trans. Softw. Eng. Methodol.*, 31(2):37e:1–37e:59, 2022. doi: 10.1145/3487043. URL https://doi.org/10.1145/3487043.

Microsoft. Microsoft's response to the European Commission's consultation on the Artificial Intelligence Act. https://blogs.microsoft.com/wp-content/uploads/prod/sites/73/2021/09/microsoft-response-to-the-european-commission-consultation-on-the-artifical-intelligence-act.pdf, 2021. [Accessed 22-08-2025].

Microsoft. How agents and copilots work with LLMs. *Microsoft Learn*, November 2024. URL https://learn.microsoft.com/en-us/dotnet/ai/conceptual/agents.

David Molnar. AI unleashed: Mastering the maze of the EU AI Act. *International University Proceedings*, 2024. doi: https://doi.org/10.56461/iup_rlrc.2024.5.ch12.

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4):2141–2168, Aug 2020. ISSN 1471-5546. doi: 10.1007/s11948-019-00165-5. URL https://doi.org/10.1007/s11948-019-00165-5.

Iohanna Nicenboim, Elisa Giaccardi, and Johan Redström. From explanations to shared understandings of AI. In *Proceedings of the DRS2022 International Conference: Bilbao*, Bilbao, Spain, June 2022. Design Research Society. URL https://dl.designresearchsociety.org/cgi/viewcontent.cgi?article=3091&context=drs-conference-papers. DRS Biennial Conference Series.

Henrik Nolte, Miriam Rateike, and Michele Finck. Robustness and cybersecurity in the EU Artificial Intelligence Act. 2024. URL https://blog.genlaw.org/pdfs/genlaw_icml2024/4.pdf.

World Trade Organization. Trading with intelligence: How AI shapes and is shaped by international trade. Report, World Trade Organization, nov 2024. URL https://www.wto.org/english/res_e/booksp_e/trading_with_intelligence_e.pdf. Comprehensive WTO Secretariat report on artificial intelligence and international trade.

Cailean Osborne, Jennifer Ding, and Hannah Rose Kirk. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. *CoRR*, abs/2405.13058, 2024. doi: 10.48550/ARXIV.2405.13058. URL https://doi.org/10.48550/arXiv.2405.13058.

Thomas O'Reilly. The EU's approach to AI is an embarrassment. *The Critic*, February 2025. URL https://thecritic.co.uk/the-eus-approach-to-ai-is-an-embarrassment/. Published in the "Artillery Row" section.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.165. URL https://doi.org/10.18653/v1/2022.findings-acl.165.

David Piorkowski, John T. Richards, and Michael Hind. Evaluating a methodology for increasing AI transparency: A case study. *CoRR*, abs/2201.13224, 2022. URL https://arxiv.org/abs/2201.13224.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023. URL https://arxiv.org/abs/2310.03693.

Tianrui Qin, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. APBench: A unified benchmark for availability poisoning attacks and defenses. *CoRR*, abs/2308.03258, 2023. doi: 10.48550/ARXIV.2308.03258. URL https://doi.org/10.48550/arXiv.2308.03258.

Elizabeth M. Renieris, David Kiron, and Steven Mills. Building robust RAI programs as third-party AI tools proliferate. *MIT Sloan Manage. Rev*, 2023. URL https://sloanreview.mit.edu/projects/building-robust-rai-programs-as-third-party-ai-tools-proliferate/.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical AI governance, 2024a. URL https://arxiv.org/abs/2407.14981.

Anka Reuel, Lisa Soder, Ben Bucknall, and Trond Arne Undheim. Position paper: Technical research and talent is needed for effective AI governance, 2024b. URL https://arxiv.org/abs/2406.06987.

Reuters. Openai rolls out cheapest ChatGPT plan at $4.6 in india to chase growth. *Reuters*, August 2025. URL https://www.reuters.com/world/india/openai-rolls-out-cheapest-chatgpt-plan-46-india-chase-growth-2025-08-19/. Updated August 19, 2025.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. TPTU: Large language model-based AI agents for task planning and tool usage, 2023. URL https://arxiv.org/abs/2308.03427.

Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A. Calvo, and Stephen Cave. Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & SOCIETY*, Feb 2024. ISSN 1435-5655. doi: 10.1007/s00146-024-01880-9. URL https://doi.org/10.1007/s00146-024-01880-9.

Waddah Saeed and Christian Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110273. URL https://www.sciencedirect.com/science/article/pii/S0950705123000230.

Christoph Schnabl, Daniel Hugenroth, Bill Marino, and Alastair R. Beresford. Attestable audits: Verifiable AI safety benchmarks using trusted execution environments, 2025. URL https://arxiv.org/abs/2506.23706.

Bruce Schneier and Nathan Sanders. The A.I. wars have three factions, and they all crave power. *The New York Times*, September 2023. URL https://www.nytimes.com/2023/09/28/opinion/ai-regulation-power.html.

Tao Shen, Didi Zhu, Ziyu Zhao, Zexi Li, Chao Wu, and Fei Wu. Will LLMs scaling hit the wall? breaking barriers via distributed resources on massive edge devices, 2025. URL https://arxiv.org/abs/2503.08223.

Mona Sloane and Elena Wüllhorst. A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. *Data & Policy*, 7:e11, 2025.

Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex 'Sandy' Pentland. Verifiable evaluations of machine learning models using ZkSNARKs. *CoRR*, abs/2402.02675, 2024. doi: 10.48550/ARXIV.2402.02675. URL https://doi.org/10.48550/arXiv.2402.02675.

F. Sovrano, E. Hine, S. Anzolut, et al. Simplifying software compliance: AI technologies in drafting technical documentation for the AI act. *Empirical Software Engineering*, 30(91), 2025. doi: 10.1007/s10664-025-10645-x.

Francesco Sovrano and Fabio Vitali. An objective metric for explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems*, 278:110866, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110866. URL https://www.sciencedirect.com/science/article/pii/S0950705123006160.

Arthur Sullivan. Europe's AI bosses sound warning on soaring compliance costs. https://www.dw.com/en/europes-ai-bosses-sound-warning-on-soaring-compliance-costs/a-70243489, 2024. [Accessed 22-08-2025].

Haochen Sun and Hongyang Zhang. PoT: Securely proving legitimacy of training data and logic for AI regulation. In *ICML 2023 Workshop on Generative AI and Law*, 2023. URL https://blog.genlaw.org/CameraReady/22.pdf.

Dariusz Szostek. Is the traditional method of regulation (the legislative act) sufficient to regulate artificial intelligence, or should it also be regulated by an algorithmic code? *Białostockie Studia Prawnicze*, 26:43 – 60, 2021. URL https://api.semanticscholar.org/CorpusID:239476730.

Marius Take, Sascha Alpers, Christoph Becker, Clemens Schreiber, and Andreas Oberweis. Software design patterns for AI-systems. In Agnes Koschmider and Judith Michael, editors, *11th International Workshop on Enterprise Modeling and Information Systems Architectures, Kiel, Germany, May 21-22, 2021*, volume 2867 of *CEUR Workshop Proceedings*, pages 30–35. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2867/paper5.pdf.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of LLM scaling based on human-generated data, 2024. URL https://arxiv.org/abs/2211.04325.

Joseph Vithayathil and Markus Nauroth. The brave new world of artificial intelligence. *Journal of Global Information Technology Management*, 26(4):261–268, 2023. doi: 10.1080/1097198X.2023.2266972. URL https://doi.org/10.1080/1097198X.2023.2266972.

Fabian Walke, Lars Bennek, and Till J. Winkler. Artificial intelligence explainability requirements of the AI act and metrics for measuring compliance. In *Digital Responsibility: Social, Ethical, Ecological Implications of IS, 18. Internationale Tagung Wirtschaftsinformatik (WI 2023), September 18-21, 2023, Paderborn, Germany*, page 77. AISeL, 2023. URL https://aisel.aisnet.org/wi2023/77.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

Andy Ward. What I've learned: Chuck Close. *Esquire*, 2007. URL https://www.esquire.com/entertainment/interviews/a2048/esq0102-jan-close/.

Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic. Agents. Technical report, Google, Feb 2025. URL https://www.kaggle.com/whitepaper-agents. Kaggle whitepaper.

Thomas Woodside and Helen Toner. Multimodality, tool use, and autonomous agents: Large language models explained, part 3, March 2024. URL https://cset.georgetown.edu/article/multimodality-tool-use-and-autonomous-agents/. Version as of August 22, 2024.

Weiyue Wu and Shaoshan Liu. Why compliance costs of AI commercialization may be holding start-ups back. https://studentreview.hks.harvard.edu/why-compliance-costs-of-ai-commercialization-maybe-holding-start-ups-back/, 2023. [Accessed 22-10-2025].

Yixin Wu and Benjamin van Rooij. Compliance dynamism: Capturing the polynormative and situational nature of business responses to law. *Journal of Business Ethics*, 168:579–591, 2021. doi: 10.1007/s10551-019-04234-4.

Doris Xin, Stephen Macke, Litian Ma, Jialin Liu, Shuchen Song, and Aditya Parameswaran. HELIX: Holistic optimization for accelerating iterative machine learning. *Proc. VLDB Endow.*, 12(4):446–460, dec 2018. ISSN 2150-8097. doi: 10.14778/3297753.3297763. URL https://doi.org/10.14778/3297753.3297763.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(3):2150–2168, 2024. doi: 10.1109/TETCI.2024.3379240. URL https://doi.org/10.1109/TETCI.2024.3379240.

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. The shift from models to compound AI systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/, 2024. [Accessed 22-10-2025].

Bruno Zulehner. EU Artificial Intelligence Act: Regulating the use of facial recognition technologies in publicly accessible spaces. Technical report, Stanford-Vienna Transatlantic Technology Law Forum, European Union Law Working Paper No. 91, 2024. URL https://law.stanford.edu/wp-content/uploads/2024/06/EU-Law-WP-91-Zulehner.pdf. European Union Law Working Papers, edited by Siegfried Fina and Roland Vogl.