Computer Law & Security Review 60 (2026) 106234

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

The regulation of fine-tuning: Federated compliance for modified general-purpose AI models

Philipp Hacker [a,*], Matthias Holweg [b]

[a] Chair for Law and Ethics of the Digital Society, European New School of Digital Studies, European University Viadrina, Große Scharrnstraße 59, Frankfurt (Oder), 15230, Germany

[b] American Standard Companies Professor of Operations Management, Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, United Kingdom

ARTICLE INFO

Keywords:
General-purpose AI
Foundation models
Fine-tuning
Regulation
Liability

ABSTRACT

This paper addresses the regulatory and liability implications of modifying general-purpose AI (GPAI) models under the EU AI Act and related legal frameworks. We make five principal contributions to this debate. First, the analysis maps the spectrum of technical modifications to GPAI models and proposes a detailed taxonomy of these interventions and their associated compliance burdens. Second, the discussion clarifies when exactly a modifying entity qualifies as a GPAI provider under the AI Act, which significantly alters the compliance mandate. Third, we develop a novel, hybrid legal test to distinguish substantial from insubstantial modifications that combines a compute-based threshold with consequence scanning to assess the introduction or amplification of risk. Fourth, the paper examines liability under the revised Product Liability Directive (PLD) and tort law, arguing that entities substantially modifying GPAI models become "manufacturers" under the PLD and may face liability for defects. The paper aligns the concept of "substantial modification" across both regimes for legal coherence and argues for a one-to-one mapping between "new provider" (AI Act) and "new manufacturer" (PLD). Fifth, the recommendations offer concrete governance strategies for policymakers and managers that propose a federated compliance structure, based on joint testing of base and modified models, implementation of Failure Mode and Effects Analysis and consequence scanning, a new database for GPAI models and modifications, robust documentation, and adherence to voluntary codes of practice. The framework also proposes simplified compliance options for SMEs while maintaining their liability obligations. Overall, the paper aims to map out a proportionate and risk-sensitive regulatory framework for modified GPAI models that integrates technical, legal, and wider societal considerations.

1. Introduction

Foundation models, or general-purpose AI models in the context of the EU AI Act,[1] are powerful AI models that are increasingly being modified by organizations for specialized uses. Yet, the regulatory and liability frameworks governing these alterations remain inadequately defined. This ambiguity primarily results from the rapid evolution of foundational AI technologies, exemplified by the emergence and widespread adoption of systems such as ChatGPT, Gemini and Claude, which led to incomplete last-minute amendments to regulatory frameworks like the European AI Act. Consequently, these foundational models have been widely recognized as a critical area requiring clearer guidelines, potential legislative revisions, or simplifications within the existing regulatory structures [1–10].

The societal importance of addressing this regulatory gap is underscored by the pervasive use and increasing scope to customise, fine-tune, or otherwise modify commercially available general-purpose AI models and systems in areas ranging from finance and education to medicine and industrial engineering [11–14] – while the legal implications are highly uncertain. Without robust regulatory clarity, there is an escalating risk of adverse societal impacts, including threats to individual well-being, violations of fundamental rights, and broader detrimental

* Corresponding author.
E-mail address: hacker@europa-uni.de (P. Hacker).

[1] We use the terms foundation(al) model, base model, and general-purpose AI model interchangeably; for a definition and disambiguation, see Section 3.

https://doi.org/10.1016/j.clsr.2025.106234

Available online 2 December 2025

consequences for the rule of law and democratic institutions [15,16]. These risks are already manifesting in an increasing number of reported AI failures characterized by misinformation, disinformation, "hallucinations", and other harmful output or misapplications of foundational AI models [17–30]. In this paper, we examine three critical implications of the modification of foundation models:

1. **Regulatory implications of model modification:** We begin by clarifying the regulatory landscape surrounding altered foundation models. For high-risk applications, the regulatory boundaries are comparatively straightforward; however, significant ambiguity remains concerning when exactly an organization qualifies as a General-Purpose AI (GPAI) provider under the AI Act (AIA). A central point of inquiry involves defining the threshold of modification: precisely at what stage does an entity transition from a mere deployer (in essence: professional user) to a provider within regulatory terms?
2. **Liability implications of model modification:** Secondly, we address the often-overlooked distinction between regulatory assessment and liability. While regulatory compliance involves adherence to defined standards, the legal liability implications arising from model modifications may differ considerably. Our analysis delineates this difference and explores its ramifications for organizations modifying foundation models.
3. **Policy and managerial implications:** Finally, we consider the broader implications from societal, policy, and practical perspectives. While regulatory frameworks can address certain societal risks of model modifications, ensuring adequate protection against individual and collective harm remains essential. Concurrently, such frameworks must balance safety with incentives for innovation and socially beneficial AI deployment. Compliance emerges as inherently multi-actor: the effective performance of modified models is always a cumulative result of both the original foundation model and subsequent alterations. We argue that compliance obligations should scale with two key factors: the extent of modification and the potential severity and likelihood of resulting harms. Additionally, we discuss voluntary codes of practice which, despite their non-mandatory nature, carry significant legal and practical consequences for adopting entities.

Our aim is to deliver a comprehensive legal analysis, formulate clear policy recommendations, and offer pragmatic guidance for organizations navigating the complexities of modifying foundation models. Specifically, we seek to identify practical approaches that achieve an optimal balance: ensuring rigorous compliance and effective risk mitigation, while simultaneously minimizing the compliance burden placed on providers.

For the purpose of our analysis, we will turn to an established framework for assessing operational risks: Failure Mode and Effects Analysis (FMEA). FMEA is a structured and systematic method widely utilized for identifying and addressing potential failure modes within processes or products [31–33]. It helps organizations prioritize risks based on their severity, occurrence, and detectability. This proactive approach enables effective risk management by facilitating early mitigation actions; this, in turn, may enhance reliability and compliance.

We utilize FMEA to systematically evaluate potential failure scenarios associated with the use of modified large language models (LLMs) and other generative AI models. Certain cases present clear-cut regulatory implications at either end of the spectrum: for example, using an off-the-shelf LLM typically places compliance responsibilities primarily on the original provider (Art. 53–55 AI Act), with limited transparency obligations for the deployer (Art. 50 AI Act), whereas developing an entirely new foundation model triggers specific regulatory requirements under the AI Act (again Art. 53–55 AI Act). In practice, however, most scenarios lie between these extremes. Organizations frequently adapt existing LLMs to their specific applications, given that creating an entirely new model is generally economically unviable. FMEA allows us to identify, assess, and manage these nuanced risks more effectively.

Some of the likely failure cases associated with modified foundation models, falling outside the clearly defined extremes, include the following illustrative examples. First, there is the case of "hallucination" resulting from fine-tuned data, where a fine-tuned LLM incorrectly accuses an individual – such as a law professor – of criminal misconduct [34].[2] Hypothetically, let us assume that this error arises from the fine-tuning dataset inadvertently containing multiple news reports of similar accusations within the same organization, which creates spurious correlations.

Second, another significant scenario involves an LLM tuned using Reinforcement Learning with Human Feedback (RLHF) [35,36], such as systems similar to Harvey [37], but providing harmful or suicidal advice during sensitive interactions. The root cause here is typically the omission of critical psychological warning signs during RLHF training, which may result in an inadequate assessment and management of risks to personal well-being [38].

A third scenario occurs when a Retrieval-Augmented Generation (RAG) LLM is deployed to support maintenance activities in industrial settings [39]. Errors arise if the system references outdated maintenance documentation, which may endanger workers due to inaccuracies. This is primarily because RAG data did not reflect recent product updates or operational changes (see the discussion in Section 3.2.4.1).

Finally, another problematic scenario involves chatbots designed to assess car insurance eligibility using historical company data (see the discussion in Section 5.2.5). Such chatbots may unintentionally discriminate against certain customer segments, as historical biases embedded in past data result in unfair or discriminatory assessments of new applicants [24,40–42].

The remainder of the paper is organized as follows: Section 2 introduces technical foundations of modifying GPAI models. Section 3 offers a detailed analysis of the ways in which these modifications may transform customizing entities from mere deployers into providers of modified models or even providers of entirely new models. Section 4 discusses the implications of modifications under the new EU product liability framework. Section 5 offers our view on the societal, policy and managerial implications of our findings, before concluding in Section 6 with a summary of the key arguments.

2. A technical perspective on GPAI modifications

A core feature of foundation models is their ability to be adjusted to specific needs by its user. This alteration comes at various levels, which adds complexity to the compliance process. The levels go from merely entering a customized prompt to fine-tuning a foundation model using reinforcement learning with human or AI feedback (RLHF and RLAIF, respectively) or both (see [43,44], and others). At the extreme, one might include distillation to build new foundation models on the interactions with an existing one.

2.1. The spectrum of modifications

For the purpose of our analysis, it is crucial to clearly distinguish among various modes of customisation when modifying foundation models. First, the simplest form of modification is using the standard architecture as provided directly by the foundation model supplier. This "off-the-shelf" scenario primarily involves alterations restricted to

[2] See also the recent case: NOYB (2025), https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer.

prompt engineering, where users interact with the unmodified model through carefully designed input prompts to achieve specific outputs.

Second, organisations may modify key hyperparameters without altering the fundamental architecture of the model itself. Hyperparameters, which are set prior to model deployment, critically influence the behaviour of predictions. An essential hyperparameter is "temperature," which modulates the randomness and creativity in model outputs. Adjusting temperature effectively balances between "exploration" (generating diverse and novel responses) and "exploitation" (favouring precise and conservative answers) [43].

Third, another prevalent approach involves RAG, where the base model architecture remains unchanged, but predictions are explicitly based on a limited, organization-specific set of documents or data repositories. Here, the model leverages external information retrieval systems, thus enhancing prediction relevance and specificity based on controlled data inputs [45,46].

Fourth, the creation of custom GPT models typically combines sophisticated prompt engineering techniques with RAG methodologies. This customization frequently utilizes specialized tools integrated within foundation model platforms, such as OpenAI's Custom GPT builder. These approaches enable users to efficiently tailor general-purpose models to specific domain applications through intuitive interfaces.

Fifth, fine-tuning methods represent deeper and more extensive customizations. These include training the model on further, domain-specific data via Supervised Fine-Tuning (SFT), as well as RLHF or RLAIF, where new data is systematically introduced into training to align model outputs more closely with human preferences and values [47]. Within fine-tuning, further specializations exist: adapter tuning involves training small, task-specific neural network modules; instruction tuning focuses explicitly on enhancing a model's ability to follow detailed instructions; and Low-Rank Adaptation (LoRA) employs efficient, parameter-saving techniques to adapt models without retraining the entire neural network [48].

Finally, distillation refers to producing streamlined versions of large foundation models, aiming for efficiency and reduced computational requirements [49–51]. In distillation, a smaller, more efficient model learns to mimic the performance of a larger, highly performant model, enabling easier deployment in resource-constrained environments.

The following table summarizes these customisation techniques:

Table 1
Main techniques for customising general-purpose AI systems.

Level of customisation	Method	Description
LOW	Standard Architecture	Using the unmodified foundation model; customization limited to prompts.
	Hyperparameters	Adjusting model settings (e.g., "temperature") without changing architecture.
	Retrieval-Augmented Generation (RAG)	Using model outputs guided by external, customer-specific datasets.
	Custom GPTs	Combining RAG with sophisticated prompt engineering via integrated tools.
	Fine-tuning (SFT; RLHF; RLAIF)	Training models further with new data and/or instructions to align outputs with human preferences of add domain-specific knowledge. Includes adapter tuning, instruction tuning, and LoRA.
HIGH	Distillation	Creating smaller, efficient models mimicking larger models for easier deployment.

2.2. Discussion: when is a modification significant?

Determining when a modification to a GPAI model becomes "significant" is essential to understanding compliance implications (see Sections 3.2.2–3.2.4). This question is pivotal because it sets the threshold at which a change in the model's behaviour may trigger a different regulatory or ethical mandate. Particularly in collaborative arrangements–where both a foundation model provider and a downstream customiser contribute to the model's development–clarifying the degree of influence each actor has on the model's performance and decisions is fundamental. To assess the compliance status of such jointly-developed models, it is crucial to delineate their respective responsibilities and the extent of their impact on the final outputs.

To unpack the compliance implications of model customisation, we must revisit the foundational mechanics of GPAI models. These systems use deep learning artificial neural networks (DLANNs) and are currently predominantly based on transformer architectures [52,53]. The operation of these models is inherently statistical: they generate outputs by processing new data–such as prompts or observations–through a pre-trained network. Accordingly, three principal components govern the behaviour of such models, each offering a potential locus for modification and thus regulatory scrutiny:

1. **Model architecture and hyperparameters:** At the core lies the model's architecture, including structural choices such as the number of layers, the nature of connections (fully connected versus functional layers), and the final decision function (like softmax or argmax). Equally influential are hyperparameters, like temperature, which affect the randomness and determinism of output generation. Adjustments in either of these areas can meaningfully alter a model's behaviour and thus bear directly on compliance thresholds.
2. **Training and validation regime:** The training process is arguably the most decisive factor in shaping a model's capabilities [54]: The selection of training data not only directly impacts performance but also relates directly to risks like model collapse [55], bias propagation and adversarial attacks [56,57]. Additionally, the training-validation split and the criteria used in the validation phase–such as accuracy thresholds or fairness metrics–further define the model's behavioural profile. These elements are integral to compliance, as they relate to principles of fairness, accountability, and transparency.
3. **Input prompts and usage context:** Finally, the inputs provided by users–whether through direct prompts or specified parameters–can steer the model's output in dynamic ways. For instance, the infamous case of Microsoft's Tay chatbot illustrated how user feedback loops can rapidly degrade model performance [58]. Prompt history also influences contextual understanding, potentially stabilising or destabilising predictions [59]. The persistent prompt instability has been vividly demonstrated in a recent study by Mirzadeh et al. [60]: they highlight that adding redundant wording alone can lead to significant output degradation. Furthermore, user-defined customisation layers, such as "personas" in generative pretrained transformer (GPT)-style systems, exemplify how downstream usage settings can significantly modify outputs.

A key outcome of the preceding analysis is the recognition that any output generated by a GPAI system must be understood as a joint function of both the foundation model (FM) developer and the downstream customizer. This joint responsibility stems from the fact that while the FM provider supplies the base capabilities of the model, it is the user's customisations–including prompt engineering, fine-tuning, and integration into specific applications–that co-determine the model's behaviour in production.

The more extensively the FM is customised, the greater the relative impact of the provider's interventions on the final output. This gradient of influence introduces an opportunity for clarifying and streamlining

compliance obligations. Specifically, if a base version of the model is retained, it enables replication and testing of outputs to determine whether a given instance of potentially harmful or unlawful content can be traced back to the original model or is the result of subsequent customisations. For the FM provider, this capability constitutes a critical compliance safeguard: if the base model, under standard test conditions, does not exhibit problematic behaviour, then it is reasonable to attribute the fault to the providers's customisation or context-specific implementation.

This framework leads to a fundamental principle: the FM provider generally cannot be held liable for compliance risks that emerge solely from how the model is applied in a given user-defined context, except in a narrow set of cases where the provider has, e.g., failed to supply adequate instructions or warnings about known model limitations.[3] Regulatory and civil liability, in this view, is conditional on foreseeability and design scope. The role of context is thus pivotal. For instance, the compliance implications of a model generating a humorous birthday poem are qualitatively distinct from its use in offering mental health advice or medical guidance, such as in the case of applications like Woebot. The latter introduces higher-stakes consequences and thus higher regulatory scrutiny, even though the underlying AI system uses fundamentally the same architecture, and most likely there will be a significant overlap in the respective training data sets (our assumption).

To support this division of compliance obligations, testing protocols should be bifurcated accordingly (see also Section 5.2.4). First, the FM provider should test and document the behaviour of the base model in a *generic environment* using *standardised benchmarks and evaluation metrics*. This forms part of the model's "system card" (as distinct from simple model cards, cf [61])–a transparency artefact that describes performance characteristics and known risks in general-use scenarios. Second, and equally crucial, the FM user or customiser should conduct thorough testing of the adapted model in the *specific context* for which it is intended, using *context-specific metrics*. This includes validating the model's behaviour under real-world conditions, assessing risks introduced by prompt strategies, fine-tuning, or integration workflows. Only this dual-testing regime can ensure comprehensive coverage that respects both the shared and context-specific contributions to the behaviour of the customised GPAI. We shall now investigate to what extent the current legal regime actually prescribes such a federated compliance framework, both under the AI Act and AI liability rules.

3. A legal perspective on GPAI modifications under the AI Act

In this section, we first define GPAI from a legal standpoint, then discuss the legal role that an organisation modifying a GPAI system may have, especially whether the modification turns them into a new "provider" of the modified model (3.2.), or whether even an entirely new model was created (3.3.). This is crucial since GPAI obligations exist primarily for providers (Articles 50 and 53–55 AI Act) and only to a very limited extent for deployers (Article 50 AI Act). The most demanding rules, by far, are contained in Article 55 AI Act, concerning basic AI safety obligations for providers of GPAI models with systemic risk (e.g., comprehensive systemic risk assessment and mitigation; red teaming; evaluations; cybersecurity; incident reporting). In short: companies not specifically developing large foundation models are well advised to take steps to avoid falling within the scope of Article 55 AI Act.

3.1. The legal definition of general-purpose AI

Building loosely of the definition of foundation models in ([1], p. 2), a general-purpose AI model is defined in Article 3(63) AI Act as an "AI model […] that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications." Hence, the two distinguishing criteria are (a) breadth of capabilities and (b) depth of possible use cases.

3.1.1. Narrow models

As a consequence, one way to altogether avoid the obligations tied to general-purpose AI models is to develop, or modify, models such that they become "narrow models." If a tool can only be used for certain coding tasks or for illustrative images, it does not qualify (see also [62], p. 630). The AI Office seems to suggest as much, too, in its preliminary approach for the guidelines on GPAI models in the AI Act ([63], p. 5). In them, the Commission makes two important points, the first of which is only partially convincing. First, the AI Office entertains a broad understanding of what qualifies as a "narrow model." Models only capable of coding, transcribing speech to text, or upscaling images are cited as examples of narrow models lacking sufficient generality ([63], p. 5). This view is mirrored in parts of the literature ([64], p. 5).

The AI Office's approach to scoping carveouts for "narrow models" in the context of the AI Act deserves critical scrutiny from a risk-based regulatory perspective. While the Commission suggests that models limited to tasks such as coding lack the generality to fall within the scope of GPAI regulation, this interpretation overlooks the generality can also relate to tasks and domains covered by a model ([64], p. 3–4), and underestimates the downstream risks even such narrowly scoped models can generate.

Consider a model that is solely capable of generating code or assisting in code production. At first glance, this may appear to be a narrow function. However, coding is not a monolithic activity. Code underlies applications in virtually every sector: finance, healthcare, defense, critical infrastructure, education, and beyond. A model that can write code across these sectors (domains) engages with a wide range of functionalities, safety profiles, and compliance requirements (tasks). The ability to generate secure authentication systems, automate trading algorithms, or manipulate sensitive patient data, even if "only" through code, introduces high-stakes risks akin to those posed by other general-purpose AI models.

Hence, the assumption that such a model is "narrow" ignores the functional breadth embedded in the act of coding itself. In this light, generality is not simply a function of the number of media formats or modalities a model supports, but also the diversity of tasks it can perform within one format, and the variety of domains it can affect. A model that generates code for embedded medical devices and simultaneously supports applications in autonomous vehicles exhibits general-purpose capabilities by virtue of its operational span, in our view.

This finding is buttressed by a law and economics perspective. The least-cost avoider principle (see, e.g., [65], p. 29) justifies upstream obligations for GPAI developers (typically one entity for one model), rather than reliance on fragmented and reactive oversight by numerous downstream implementers [4,66,67]. The risks introduced by a "coding-only" model do not remain contained. They propagate through the AI value chain as third parties build layered applications upon it, often without full visibility into the model's design, limitations, or training data. This diffusion of responsibility increases risk and reduces accountability. Regulating such models as GPAI aligns with the efficiency principle of risk containment at the source, where the developer has the greatest technical capacity, and generally the cheapest economic opportunity, to implement safeguards. GPAI coverage means: Only one entity has to act at the source, and not a thousand entities downstream.

Therefore, the Commission's first point, which frames models limited to coding or text-to-speech functions as inherently narrow and outside GPAI scope, appears conceptually and economically weak. A functional and risk-aware understanding of generality should inform the regulatory classification, rather than a view of modality or use-case singularity. Hence, we would restrict narrow models to those models

[3] There are further exemptions to this general rule considered in detail below, see Section 4 and note 6 et seq. (e.g., foreseeable misuse).

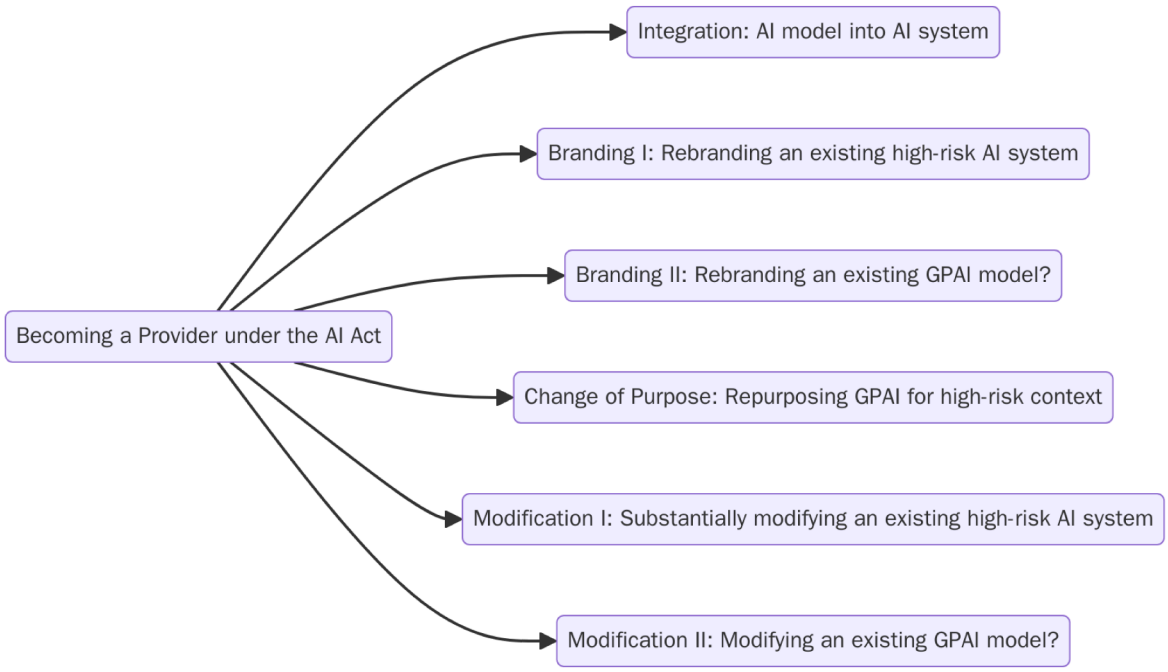

Fig. 1. Overview of the six paths to transitioning from a deployer to a provider.

only executing some specific tasks within a defined purpose or modality: in coding, for example, only syntax correction, code formatting, or autocompletion within a predefined programming language and domain; in image generation, only upscaling; in language generation, only completion, or summary, etc.

3.1.2. *FLOP thresholds as a proxy*

Second, the AI Office suggests a threshold of 10^{22} floating operation points (FLOPs) above which a presumption is triggered in favour of, and below which against, sufficient generality ([63], p. 4). These presumptions are supposed to be rebuttable. Quite obviously, they are modelled on the 10^{25} FLOP presumption in Article 51(2) AI Act concerning GPAI models with systemic risk. While not explicitly foreshadowed in the Act, the new 10^{22} capability threshold coupled with rebuttable presumptions present, in our view, a decent proxy providing a bright red line for companies seeking to determine whether they are in or out of scope of GPAI regulations.

FLOP threshold are, of course, necessarily a crude instrument; but so are all alternative proxies, be they benchmark performance, model size in terms of parameters, or others [68]. Progress, particularly in model distillation and reinforcement learning, may soon make high FLOP thresholds obsolete, as the DeepSeek model family demonstrates [69]. Nonetheless, these cases could be dealt with outside of the presumptions: for example, if a model is distilled from an above-threshold model, basically inheriting its capabilities, or produced otherwise to mimic the capabilities of an above-threshold model, then it should still count as possessing the required generality and capabilities, barring evidence to the contrary. Overall, as long as no academic consensus exists on proxies for capabilities, proxies such as FLOP thresholds may be used to guide legal categorization, but must be accompanied by a description of their weaknesses and elements considered to overcome them in cases where the presumption does not match reality (anymore): Just like Annex XIII AI Act lists criteria for the designation of general-purpose AI models with systemic risk referred to in Article 51 for cases in which the FLOP presumption does not lead to desirable results. Over time, proxies and the criteria for designation outside of the proxy thresholds can be and must be adapted, of course.

3.1.3. *Defining GPAI models with systemic risk*

As mentioned, the most comprehensive rules for foundation models are reserved, in Article 55, for GPAI systems presenting systemic risks. These models are defined by the AI Act, specifically Articles 3(64) and 3 (65). The latter delineates 'systemic risk' as "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain." Art. 3(64), in turn, specifies that "'high-impact capabilities' means capabilities that match or exceed the capabilities recorded *in the most advanced* general-purpose AI models." [our emphasis]

However, this definition introduces a critical issue: by referring exclusively to the most advanced models at any given moment, there is a risk of excluding models that were previously considered cutting-edge but have since been superseded. This creates a scenario where a model's risk categorization might shift over time, potentially exempting significant models–such as GPT-4 after the introduction of GPT-4o, or Llama 3.1 (405B parameters) once succeeded by Llama 4–from systemic risk designation despite their continued relevance.

We can imagine two interpretations of this definition: a dynamic interpretation continually categorizes only the top few models (e.g., the top five chatbot models at any given time, as measured by appropriate benchmarks) as systemic risk models. The fundamental problem here is that models would continually be removed from the systemic risk list as newer models emerge, despite no actual reduction in their inherent risk profiles. Indeed, if anything, the release of new models typically results in *decreased* safety oversight and maintenance of incumbent models, potentially *heightening* their risk.

A second, and more pragmatic interpretation would treat "most advanced" models as those identified as most advanced at the time of the AI Act's enactment (August 2024), or those surpassing a defined capability threshold that remains relatively stable over time. Although this threshold may incrementally increase as technology advances, it would provide a fairly fixed anchor. This would ensure that models do not easily lose their systemic risk classification simply due to the release of newer, more powerful models. This static approach also provides greater legal certainty and maintains consistent safety standards for significant

AI models over time. From a systematic perspective, it is buttressed by the reference, in Art. 51(2) AI Act, to the fixed (even though generally adaptable) FLOP threshold for the systemic risk presumption. This fixed presumption approach in Art. 51(2) is effectively impossible to reconcile with a dynamic interpretation of Art. 3(64). Hence, we believe that regulators and courts should follow a static approach, which is also the only one compatible with the risk-based framework of the AI Act. This architecture and purpose would be thoroughly undermined if models would automatically drop from the systemic risk category even though their risk profile remains entirely unchanged.

3.2. Becoming a provider of a modified model under the AI Act

Having defined what a GPAI model is, we can now turn to the crucial question of the regulatory status of the entity engaging with it. In the context of the AI Act, the distinction between "provider" and "deployer" carries significant regulatory consequences, particularly where high-risk applications, for example in recruitment or other employment settings, or systemic risks under Article 55 are concerned. Providers bear the brunt of ex ante obligations, including technical documentation, conformity assessments, and ongoing risk management duties. By contrast, deployers face narrower, often operational, compliance requirements under the Act (Article 26–27). This asymmetry creates strong incentives for actors–especially those operating in high-risk domains such as recruitment or working with GPAI models–to avoid classification as a provider unless it is strictly necessary. The stakes are even higher where a model may fall under the systemic risk framework, which introduces heightened scrutiny and safety obligations under Article 55. As a result, firms will – and should – seek to structure their role and contractual arrangements in a manner that minimizes exposure to the provider designation and its attendant responsibilities.

3.2.1. Six paths to becoming a provider

The categories of deployer and provider, however, are not fixed; rather, deployers can transform into providers, sometimes without their knowledge. The AI Act provides six routes to becoming a provider (see Fig. 1), but they are specified in incomplete ways in the legal provisions.

The **first path** to becoming a provider under the AI Act arises if an entity integrates GPAI model into an AI system, as set out in Article 3 (68). This provision defines a "downstream provider" as "a provider of an AI system, including a general-purpose AI system, which integrates an AI model, regardless of whether the AI model is provided by themselves and vertically integrated or provided by another entity based on contractual relations." Simply integrating a GPAI model, however, does not automatically trigger the full set of provider obligations. The downstream provider becomes a system, not a model provider. This difference, between models and systems, is crucial for understanding GPAI rules in the AI Act. Systems are the interfaces interacting with users; models are the mathematical objects enabling AI systems (see Article 3(1), (66) and particularly Recital 97 AI Act[4]). Importantly, Articles 53 to 55 only apply to model providers, not system providers. In cases where the resulting system does not fall into a high-risk category or involve a prohibited use, the obligations for downstream providers are, therefore, limited to baseline requirements such as transparency measures (Art. 50) and ensuring AI literacy among staff and agents (Article 4).

The **second path** is branding, where an entity rebrands a high-risk AI system (i.e., a system already placed on the market or put into service) as its own by applying its name or trademark, e.g. logo, to the system. According to Article 25(1)(a) AI Act, this act makes the institution legally responsible as a provider, even if it did not develop the AI, or have it developed, originally. The reason for this rule is that by branding, the institution presents itself as responsible for the system, endows the system with trust in its brand and commits towards it the goodwill users and consumers may harbor. Hence, it assumes all the duties of a high-risk provider. Neat representation through branding is paid for by legal exposure.

Third, the same logic should apply, by analogy, to the branding of GPAI models (see also [70], para. 76). For example, if the hypothetical "Prompt Company" uses a version of GPT-4 or an open-source model and makes an AI system based on it available to its employees for knowledge tasks, branding it as the "PromptCompany-GPT," it would qualify as a provider of the GPAI model (and system) under this prong (Art. 25(1)(a) AI Act by analogy). The use of trust and goodwill in the name or brand is exactly the same in cases of GPAI models as in high-risk systems. Hence, branding entities should be responsible for complying with the main rules governing these AI tools – the GPAI model rules in this case, just like the high-risk system rules in the other case. This transition does not, however, overburden deployer since it is very clear to apply and easy to avoid. Simply put, they must not associate any GPAI model with their own name or trademark in ways to suggest a stake in its development. It should be sufficient to use a fantasy name, such as "UnicornGPT", to avoid this path.

The **fourth path** to becoming a provider is a change of purpose. More specifically, the intended purpose of an AI system already placed on the market or put into service must be changed from non-high-risk to high-risk (Annex I Section A or Annex III). If an institution starts using a non-high-risk AI system including a GPAI system in a high-risk context, such as employing it for tasks like credit scoring, life or health insurance, or recruitment decisions, it triggers Article 25(1)(c) AI Act. For example, if the HR department of a company decides to run a few candidates through ChatGPT for initial screening, the company immediately becomes a provider of a high-risk AI system – even if the result is hand-checked and the remaining recruitment process does not use AI at all. By changing the purpose of the GPAI to a high-risk use, the deployer assumes the role of a provider and must ensure that the AI system complies with the high-risk AI regulations under the AI Act [4]. Again, this transition is clearly defined and easy for companies to avoid if all persons dealing with AI systems know the high-risk areas – which they are legally required to under the AI literacy obligations of Article 4 AI Act.

The most difficult and interesting cases are presented by the remaining two alternatives, precisely surrounding fine-tuning and other modifications. The **fifth path** is through substantial modification of a high-risk AI system (which is already placed on the market or put into service), covered by Article 25(1)(b) AI Act. For example, if the institution makes significant changes to such an AI system, such as fine-tuning it with proprietary data or altering its core functionality, it may be re-classified as a provider. A substantial modification is defined as a change that was not covered by the original conformity assessment and that either affects the AI system's compliance with the requirements in Articles 8–15 or changes its intended purpose (Article 3(23) AI Act). Even modifications like fine-tuning with new data sources or additional training can be considered substantial if they significantly alter the system's behaviour or risk profile. We shall investigate this in detail belowSections 3.2.3.

The **sixth and final path** is very similar to the fifth one, but starts with a GPAI model at the outset. This – finally – covers the classical industry use case of picking a GPAI model, customizing it in some way by means of modifications, and then putting it to use, either internally or externally. Importantly, if GPAI models are used without fine-tuning or other modification by institutions, the institutions themselves remain mere deployers and do not become providers (unless one of the first four cases occurs). However, if a GPAI model (e.g., GPT-4o; Llama 3.1 70B; Mixtral 8 × 22B) is fine-tuned or modified by the institution, the entity

[4] See Recital 97, sentences 5-7: „Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems. AI models are typically integrated into and form part of AI systems."

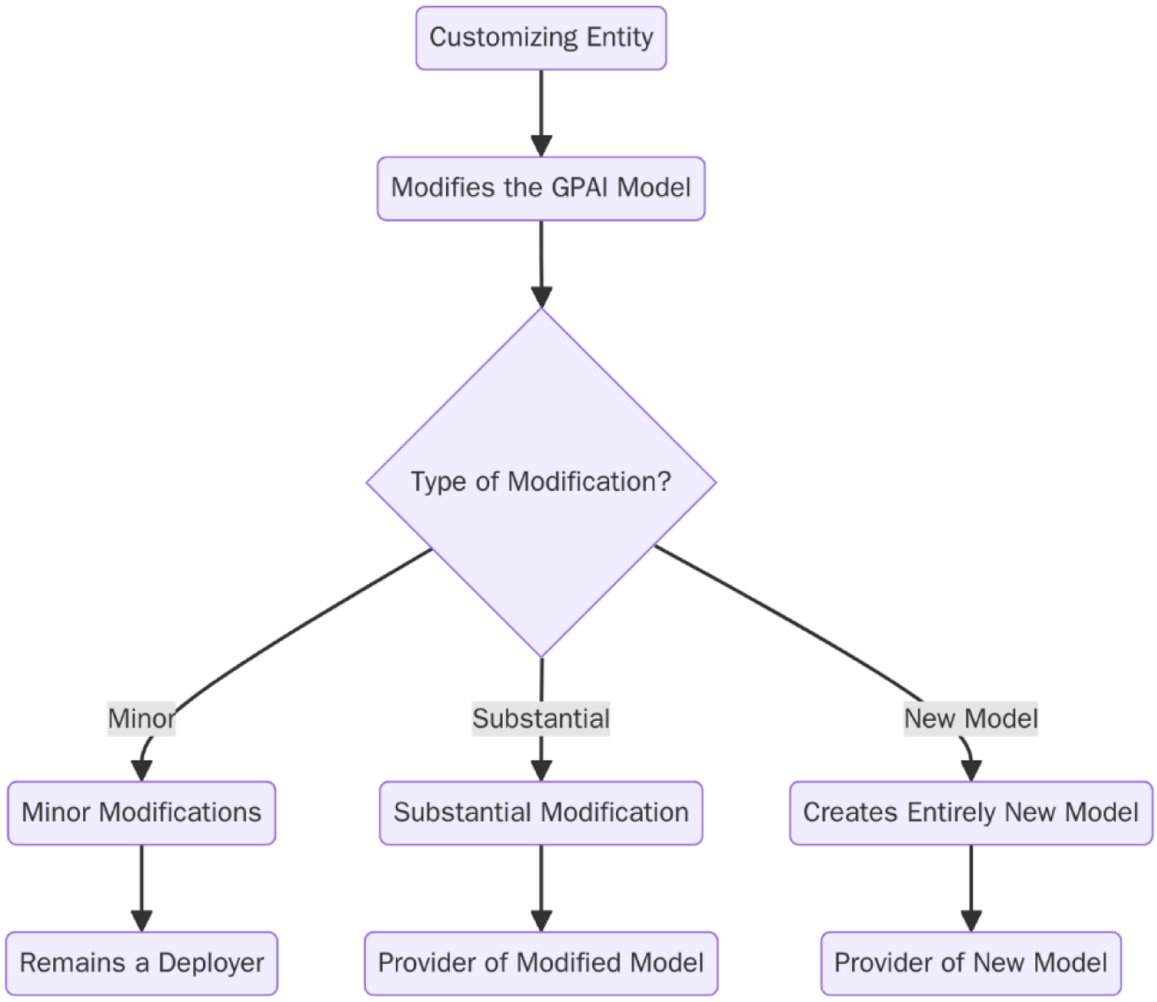

Fig. 2. Overview of legal consequences for fine-tuning or otherwise modifying a GPAI model.

may become a provider with respect to the changes it made to the model, as suggested by Recital 109 AI Act. While this is the practically most relevant scenario, there is, unfortunately, least clarity on the consequences concerning this case in the AI Act.

If a deployer modifies or fine-tunes a GPAI model, there are two options for interpretation: first, one could argue that the AI Act does not cover this scenario at all (unless Art. 25(1)(c) applies), and the entity conducting the fine-tuning does not become a provider. After all, Article 25 only contemplates the modification of high-risk systems, not of GPAI models. Second, alternatively, the solution may be found in Recital 109, which holds: "In the case of a modification or fine-tuning of a model, the obligations for providers of general-purpose AI models should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources, as a means to comply with the value chain obligations provided in this Regulation." That part can, again, be read in two ways. It can mean that only "original" providers of GPAI models have (limited) obligations in case of modification/fine-tuning; or it can mean that (any) entity conducting the modification/fine-tuning may become a provider of the modified/fine-tuned model.

While neither interpretation can be ruled out, arguments are much stronger for the latter. First, it does not make sense to impose obligations in the case of modification or fine-tuning on the "original" provider (e.g., OpenAI in the case of GPT-4o; Meta in the case of Llama 3.1 70 B) as that original provider often will not – and must not – know of the fine-tuning activity. Second, systematically, the recital specifically references the value chain obligations, i.e. Art. 25, and hence the switch-to-provider rules. If only the original provider was covered, the reference would be obsolete, as there is no legal value chain in that case. Third, from the perspective of the purpose of Art. 25 and Rec. 109, entities engaging in (certain types of) fine-tuning need to be included. Otherwise, an organization could use a GPAI model with systemic risk and fine-tune it on hate speech (e.g., 4chan), creating a much riskier hate bot. However, that organization would not be bound by Art. 55 and its risk mitigation rules if obligations under the AI Act were limited to original providers. Neither would that original provider, however, be covered as it has not conducted the fine-tuning. Art. 26 AI Act does not apply, either, as a chatbot is not a high-risk system, even if it spews mainly hate. Hence, no entity would be obliged to perform any of the AI safety evaluations and mitigation practices that the AI Act typically requires in such cases. This does not seem to be a convincing result that protects fundamental rights. Hence, while clarity can only be provided by a CJEU judgment, it is reasonable to assume that certain types of fine-tuning and modifications by deployers transitions them into providers, based on an analogy to Art. 25(1), interpreted in the light of Recital 109. Institutions may want to err on the side of caution by assuming that at least significantly fine-tuning or otherwise modifying a GPAI model makes them a provider of that new model, as we discuss below.

Whether or not GPAI modifications lead to a provider status depends, in our view, on whether this comprises (Fig. 2)

- a substantial modification (Art. 25(1) by analogy; see under 3.2.2.–3.2.3.),

- a minor, insubstantial modification (Art. 25(1) by analogy; see under 3.2.4.), or even
- an entirely new model (Rec. 97; see under 3.3.).

3.2.2. Differentiating substantial from insubstantial modifications

Under the AI Act, the classification of a modification to a GPAI model as substantial or insubstantial must, arguably, follow a risk-based assessment. This approach aligns with the fifth regulatory path described above, under Article 25(1)(b) and Article 3(23) of the AI Act, which defines substantial modifications in terms of their potential to alter the system's compliance or risk profile.

Ideally, hence, the central criterion is whether the modification leads to a substantial change in the risk profile of the model, particularly with regard to fundamental rights or other protected legal interests. A modification should be considered substantial when it substantially increases at least one relevant risk, such as the risk of discrimination, the generation of inaccurate or harmful content (often referred to as "hallucinations"), or biochemical or even nuclear risks. By contrast, a modification should be treated as insubstantial when it merely enhances the model's behaviour across several risk dimensions and worsens it in only one, provided that the improved risk dimensions are the most relevant to the intended purpose of the system. Only this reading, in our view, offers a purpose-sensitive interpretation of "substantiality," and is rooted in the AI Act's broader commitment to proportionality.

A challenge arises, however, when one seek to apply this standard to GPAI systems. In practice, assessing changes in the risk profile requires statistical testing. This assessment must be performed on a case-by-case basis, which can be particularly onerous for downstream deployers who do not develop the base model but only fine-tune it. It may be tempting to assume that fine-tuning on a small dataset constitutes an insubstantial modification by default. This interpretation would allow many downstream companies and other users to avoid the burdens of independent testing. However, this assumption is technically incorrect.

Even minimal fine-tuning can induce emergent misalignment, where the model's behaviour changes significantly and unexpectedly [15]. Additionally, attacks such as model poisoning or malicious injection form part of a number of techniques where small interventions can have major negative effects [71–73]. This undermines any categorical distinction based solely on formal criteria like dataset size, parameter count, or the volume of modified data.

To address this challenge, three regulatory approaches may be considered.

1. **Risk-Based Approach:** Under the first approach, each instance of fine-tuning is treated as a potentially substantial modification unless comprehensive evaluations demonstrate that the model's risk profile remains materially unchanged. This interpretation is consistent with the underlying risk-based logic of the AI Act and offers a principled way to account for the unpredictable effects of model adjustments, such as emergent misalignment. Implementation of this approach would require deployers or downstream providers to conduct structured and rigorous testing before the deployment of a fine-tuned model. These evaluations must assess whether the modification has introduced new risks or exacerbated existing ones. The process must be documented in sufficient detail to enable both internal governance and external regulatory scrutiny. However, while this approach respects the risk-based character of the AI Act, it imposes considerable burdens on downstream actors who may lack the technical or institutional capacity to carry out such evaluations systematically.
2. **Innovation/Simplification-Focused Approach:** A second approach, oriented more toward regulatory simplicity and innovation support, begins from the opposite assumption. It treats minor modifications–including fine-tuning on small datasets–as insubstantial by default, even though such changes may sometimes alter the model's risk profile significantly. The justification here lies in the need to facilitate downstream use of GPAI systems without imposing excessive compliance burdens. This approach appears to be favoured by the AI Office. In its consultation document, the Office proposes the introduction of a compute-based (FLOP) threshold ([63], p. 9). Under this model, an entity would only be treated as a new provider if it uses at least one-third-of the FLOPs originally required to develop the model. In the case of systemic-risk GPAI models, the same rule applies with an additional alternative: a deployer would also become a provider only if the combined compute of the original model and the modification first crosses the systemic-risk threshold of 10^{25} FLOPs. This prong rightly prevents entities from circumventing the systemic-risk obligations by fragmenting the development process across multiple actors. The appeal of this approach lies in its clarity and operational simplicity. The FLOP threshold functions as a bright-line rule and makes compliance more predictable and enforceable. Moreover, the limitation of Article 25 obligations to high-risk AI systems suggests that the legislators intended to focus GPAI duties primarily on original providers, rather than extending them comprehensively to downstream deployers. However, Recital 109, to a certain extent, speaks against this approach. There, the legislators explicitly acknowledge that the value chain responsibilities set out in Article 25 are incomplete, and that additional obligations must sometimes apply to downstream entities. This implies that treating minor modifications as generally insubstantial may overlook legally relevant risks. Moreover, some downstream modifications do not correlate at all with compute-based thresholds, such as when Retrieval-Augmented Generation (RAG), filters, or system prompts are used.
3. **Our Approach: Compute and Consequence Scanning (CCS):** A third and, in our view, preferable approach combines the regulatory clarity of the FLOP threshold, in areas where it does apply (essentially, fine-tuning or re-training) with a functional yet simple assessment of the risks introduced by model modification. FLOP-based metrics alone are insufficient, not only because they do not apply to all modifications, but also because even minor interventions in terms of compute can lead to major behavioral shifts, as discussed. Instead, the FLOP threshold should serve as a baseline rule, always to be supplemented by a structured process of consequence scanning drawn from the FMEA toolkit. This amounts to a compute and consequence scanning test (CCS test): combining the bright red line of a compute threshold with the moderate burden of functional risk analysis. Under this approach, downstream entities remain subject to the FLOP-based test proposed by the AI Office. One may always argue about the right threshold limit, and one-third-may initially seem too high. Only very few, heavily resource-intensive modifications would cross that threshold. It does, however, seem justifiable to pick a high threshold if it is *always* combined with consequence scanning. Thus, if a technically informed consequence scan indicates the presence of major risks resulting from the modification, then the entity should be treated as a provider – even if the FLOP threshold of one-third-is not exceeded. This hybrid test enables a more calibrated regulatory response. For instance, a RAG system that retrieves documents from an ordinary corporate archive typically does not raise significant new risks and would not trigger provider status. By contrast, a RAG system that retrieves documents from an atomic oversight agency would likely surface major risks under consequence scanning. Similarly, fine-tuning using proprietary, well-curated industrial data generally does not introduce major concerns, while fine-tuning using data from a source known to contain significant bias would or hate speech. Furthermore, even where entities remain developers and thus fall outside the scope of certain AI Act provisions, they may still bear legal responsibilities under adjacent legal regimes. In particular, under the revised Product Liability Directive (PLD) and general tort law, such entities could qualify as manufacturers or users of a dangerous product and thus be liable for certain types of harm caused by the modified AI system, regardless of

whether they are considered providers under the AI Act, as we investigate below (Section 4).

Overall, the test for distinguishing substantial from insubstantial modification must ultimately turn on functional risk, not formal proxies. A composite approach that integrates FLOP-based rules with consequence scanning offers the most balanced response, in our view, by preserving legal coherence within a risk-based framework while enabling both innovation and accountability.

3.2.3. Insubstantial modifications

Applying the CCS test outlined in the previous section to concrete cases illustrates which types of modifications typically qualify as insubstantial under the AI Act. These are interventions that do not alter the core architecture, intended purpose, or risk profile of the underlying AI system. In particular, minor technical adjustments–such as changing hyperparameters or applying post-processing filters–will, in most instances, fall below the FLOP threshold of significance. Nevertheless, consequence scanning remains necessary in each case to ensure that no material risk is introduced, even where the modification appears minor on its face.

3.2.3.1. Retrieval-augmented generation. RAG refers to the practice of augmenting an AI model's outputs by integrating external data sources retrieved at runtime [46]. This enhancement changes the context in which the model generates responses but does not alter the model's internal weights, architecture, or training data. As a case in point, remember the third scenario from the introduction, in which a RAG system is deployed to the support maintenance in a machinery context, but references an outdated database with erroneous system specifications, which in turn leads to harm to workers.

Because RAG operates at the interface level and leaves the underlying model untouched, it generally qualifies as an insubstantial modification. The model's core functionality and intended purpose remain intact. The flop threshold does not apply here as RAG does not necessarily, and not even typically, involve compute spent on retraining the model.

However, the consequences of the retrieval operation must still be assessed. If the documents retrieved involve highly sensitive domains – such as intelligence, nuclear oversight, or biomedical attack response – the consequences of using those sources may trigger material risks. Thus, even in cases where the architecture remains unchanged, consequence scanning is necessary to confirm the insubstantial nature of the modification.

The scenario of maintenance in an industry context presents a good test case for CCS, and the considerations in applying it. On the one hand, consequence screening may reveal, if heavy machinery is involved, that serious adverse consequences may occur if the RAG system does not function as planned; for example, workers might be hurt or killed. On the other hand, even the tests required under Article 55 AI Act are not necessarily well-suited to detect errors such as outdated information in RAG databases. Rather, the evaluations required under Article 55 would often compare the RAG system output against the "ground truth" contained in the database. If the database was initially wrongly chosen (contains erroneous data), such a test would not reveal the error.

Rather, what is needed are incentives to choose the right database with the correct records. This, however, is a problem unrelated to the behavior of the LLM itself. Put differently, the problem arises at the system, not the model level. Hence, what is needed are incentives to correctly build the *system*, without breaching a standard of care. These incentives are provided by the tort law system (and the high-risk rules of the AI Act, if they apply), not by GPAI model regulation. Indeed, the organization developing the RAG system would quite clearly be liable under tort law for any harm occurring to workers (see Section 4.2). Hence, we submit, only in cases where the tort law incentives need to be buttressed by specific GPAI regulation should a new provider status be assumed. This only holds in cases of truly catastrophic consequences arising from the specific RAG context (e.g., ABC risks for RAG systems used in nuclear power plants or biochemical attack response).[5] In all other cases, such as the industrial maintenance application, consequence scanning does not generally reveal risks that would warrant the application of a new *model* provider status. Rather, the tort law system and the high-risk rules at the system level (cf. Art. 25(1)(c) AI Act) deal with such risks. Overall, RAG systems will, thus, generally not trigger GPAI model provider duties under the AI Act.

3.2.3.2. Fine-tuning with minor adjustments. Fine-tuning with minor adjustments involves training the model on a limited dataset for a short duration to improve performance on specific tasks. Such fine-tuning may involve light-touch adaptation of the model's parameters without altering its core behaviour, domain focus, or compliance characteristics.

These types of interventions often do not affect the system's conformity with the AI Act or its fundamental use case. Accordingly, they are generally treated as insubstantial modifications under the "one-third-FLOP threshold". However, this conclusion cannot rest solely on the limited scope of the modification. As mentioned, empirical research show that even small fine-tuning operations can trigger emergent misalignments and significantly more harmful outcomes, especially if the fine-tuning data introduces bias, contradicts safety constraints, or alters model behavior in edge cases. Because of this risk, consequence scanning remains necessary.

For example, fine-tuning a model with a dataset that contains skewed demographic data may inadvertently reinforce or introduce discriminatory outputs in high-stakes contexts such as housing or car insurance (both not high-risk use cases). Similarly, using a limited set of adversarial or fringe-content examples to fine-tune a chatbot could cause it to generate conspiracy theories or manipulative content more reliably, particularly if the model overfits to this new content (see, for the opposite direction of tuning a model for more aligned output using limited data, [74]). A further example would be a model fine-tuned for internal use by a law enforcement agency, where the added data may shift its behavior in ways that affect due process or fundamental rights protections, even if the fine-tuning was modest in scale. In such cases, despite the limited compute and dataset size, the downstream risk implications warrant treatment as a substantial modification. This, in turn, triggers the provider obligations under the AI Act.

3.2.3.3. System prompts. System prompts involve altering the instructions that guide the model's behaviour during inference without modifying the model's architecture or training data. This approach affects only the prompt engineering layer and does not alter the underlying model's structure, performance characteristics, or intended purpose.

Again, the FLOP threshold is irrelevant here. As a rule, introductions of or changes to system prompts constitute insubstantial modifications under the AI Act. However, there are important exceptions. If the prompt itself introduces or amplifies a major risk–such as an instruction to produce hate speech, deceptive output, or privacy-violating content–then the modification may become substantial due to its downstream impact. These cases do not involve architectural changes but nonetheless trigger specific behaviour and risk exposure that requires regulatory attention via provider status. Consequence scanning is required to draw this line.

3.2.3.4. CustomGPTs. CustomGPTs operate in a manner similar to system prompts. They are created by combining existing foundation models

[5] Note that these cases will likely qualify as high-risk cases and trigger system provider duties under Art. 25(1)(c) AI Act, anyways.

with customized instructions, memory modules, or conversation settings tailored to specific user needs. As with system prompts, these changes do not modify the underlying model's weights or architecture and typically preserve its intended function.

Therefore, the FLOP threshold does not apply, and CustomGPTs are normally regarded as insubstantial modifications. Nonetheless, the same caveat applies: if the customization substantially increases risk–especially where the memory module or prompt chain is designed to elicit harmful, misleading, or discriminatory outputs–then the modification may become substantial. For example, a CustomGPT designed to impersonate legal or medical professionals without safeguards could expose users to serious harm. Similarly, a CustomGPT constructed to simulate negotiation tactics that intentionally deceive may raise concerns. In each case, consequence scanning may lead to the conclusion that it triggers the provider obligations under the AI Act. Note, however, that the high-risk rules for AI systems are independent of this assessment. Hence, a CustomGPT parsing and evaluating CVs efficiently for recruitment would (likely) not lead to a model provider status, but to a high-risk system provider status via Art. 25(1)(c) AI Act.

3.2.4. Substantial modifications

A modification to a GPAI model qualifies as substantial under the AI Act when it introduces a new or substantially negatively altered risk profile, particularly in relation to fundamental rights, safety, or other legally protected interests. In such cases, the entity making the modification may be treated as a new provider under Articles 25(1)(b) and 3 (23) by analogy and must assume the full set of GPAI provider obligations.

As mentioned, substantiality generally presupposes that the internal parameters of the model–its weights and biases–are changed, because these alterations typically affect how the model functions in practice. The FLOP threshold proposed by the AI Office, which defines substantial modification as one involving at least one-third-of the compute originally used to train the model, functions as an important–but not exhaustive–proxy. As mentioned, the threshold cannot be used as a substitute for consequence scanning, which remains necessary to capture modifications that materially alter the model's behaviour even if they fall below the FLOP threshold.

Significantly, in our view, the potentially changed behaviour needs to relate to systemic risks for the modifying entity to fall under Article 55 (see Section 3.1.3), while a significant change of any risk relevant under the AI Act is sufficient to trigger the obligations under Articles 53 to 54 AI Act (cf [63], p. 9–10).

3.2.4.1. Fine-tuning with substantial changes. Fine-tuning becomes a substantial modification when it involves retraining the model on large datasets, substantial parameter updates, or task reorientation that meaningfully changes how the model behaves. This includes domain shifts (e.g., medical GPAI to legal GPAI contexts) or optimization for different user groups or applications.

Where such fine-tuning involves compute resources above one-third-of the FLOPs used in the original model, the AI Office proposal would presumptively treat the modifying entity as a provider. Our CCS test of significance applies both quantitatively (via FLOPs) and qualitatively (via consequence scanning), with either being sufficient to trigger provider obligations, unless extensive evidence to the contrary is produced.

Remember the first case from the introduction, in which the fine-tuning data set contained news reports which led to accusations of criminal misconduct against an individual in the output of the modified model. Here, the consequence scanning would indeed have revealed a significant risk of "hallucinations" violating personality rights (and potentially copyright) stemming from the content of the fine-tuning data set: news reports including personal data. Hence, even if the one-third-FLOP threshold was not reached, the fine-tuning leads to a substantial modification, at least concerning general risk. This may be different if fine-tuning involves more "innocuous data," such as general industry or machinery data.

More generally, the classification as "substantial" will depend on the propensity of the fine-tuning data to trigger general (Articles 53–54 AI Act) or even systemic model risk (Article 55). As mentioned, the law must distinguish between a substantial modification concerning general risk, triggering provider status under Articles 53 and 54, and concerning systemic risk, additionally triggering provider status under Article 55 AI Act. Whether "hallucinations" indeed amount to systemic risk under Article 3(65) AI Act cannot be investigated in depth in this piece; however, this appears at least possible (see Section 3.1.3).

3.2.4.2. Reinforcement learning with human or AI feedback (RLHF/RLAIF). Reinforcement Learning with Human Feedback (RLHF) and its variant with AI Feedback (RLAIF) involve post-training optimization where the model's outputs are ranked and weighted according to a preference function [36,75–77]. These techniques update the model's internal weights to better align its behaviour with certain objectives, often by prioritizing safety, helpfulness, or adherence to ethical guidelines [25,77]. As an example, let us revisit the second case from the introduction, in which a chatbot model was tuned with RLHF, but psychological warning signs were omitted. Hence, the model occasionally provides suicidal advice, as it apparently happened in the case involving character.AI [78].

These processes typically require substantial compute resources and may even exceed the one-third FLOP threshold, particularly when applied across wide model capabilities; for OpenAI's o1, post-training overall may have accounted for 40 % of overall compute [75]. Even when compute usage is below this formal limit, RLHF or RLAIF can significantly reshape how the model makes decisions or navigates trade-offs between competing values. In consequence, these forms of fine-tuning generally result in a substantial modification. They require full consequence scanning, and typically further testing evidence, to determine whether the risk profile remains stable. Absent such validation, the modifying entity must be treated as a new provider under the AI Act. In the above case of chatbot tuning with RLHF without psychological triggers, the risk profile may change such that, typically, the underlying model is substantially modified and the customizing entity becomes a new provider, as evidenced by potentially harmful advice in psychologically sensitive situations.

3.2.4.3. Jailbreaking and related interventions. Jailbreaking includes any intervention that seeks to circumvent or dismantle alignment constraints, safety filters, or access controls embedded in the model [79,80]. This may involve modifying system-level configurations, injecting adversarial prompts, or altering the model's internal logic to enable previously restricted outputs [72,73].

If these interventions are accomplished by modifying internal parameters–for instance, by changing or bypassing weights responsible for filtering or refusal behaviour–then they typically qualify as substantial modifications. While such jailbreaks will often not involve compute resources that meet the FLOP threshold, they often do lead to disproportionate changes in behaviour, and expose users or third parties to heightened risks. In these cases, the absence of high compute usage does not exempt the actor from provider responsibilities. If consequence scanning reveals any major risk increase, significance must be presumed and regulatory obligations attach accordingly.

Overall, while the FLOP threshold of one-third-offers a useful starting point for identifying substantial modifications, it cannot serve as the sole criterion. Risk must also be assessed functionally. Any change that significantly affects the model's risk profile–whether or not it crosses the FLOP threshold–should be treated as substantial. The dual-track test of compute and consequence scanning (CCS) ensures that both overt and subtle risk shifts are captured by the regulatory framework.

3.3. Creation of entirely new models under the AI Act

The development of an entirely new AI model, as distinct from the modification of an existing one, carries a different set of regulatory implications under the EU AI Act. Most importantly, Article 25 does not apply to the creators of entirely new models. As a result, the specific obligations related to collaboration across the AI value chain, as set out in Article 25(2), are not triggered. Instead, the entity developing a new model is subject to the full scope of provider obligations under the Act "from scratch", as it is considered the primary source of the system's design and development.

3.3.1. Defining a new model

In our view, an AI model is considered entirely new when it either incorporates substantial architectural changes or exhibits very significantly different behaviour from any pre-existing model. This distinction can be based on structural, functional, or operational elements.

From an architectural perspective, a model built from scratch, particularly but not only when using a novel framework, training setup, or learning paradigm, qualifies as new. For example, models that diverge from the transformer-based architectures predominant in current LLMs fall clearly into this category. However, it would also be sufficient that Company B trains a new model from scratch, entirely copying the architectural choices of a model developed by Company A.

From a behavioural perspective, a model that operates in ways that differ qualitatively from its predecessors may also be considered new, even if it is based on similar architectural foundations and was developed by the same company as a similar model. The difficulty here lies in distinguishing mere different model versions from new models. A simple improvement in performance–such as faster inference or higher benchmark scores–does not generally suffice for a new model. However, if the model demonstrates qualitative behavioural differences on standard evaluation sets (e.g., shifts in "reasoning" strategy, alignment characteristics, or risk profiles), this may support the classification as a new model. A change in version number, such as the progression from GPT-3 to GPT-3.5, can serve as an indication of this shift, although it is not a determinative criterion on its own.

The AI Office suggests using the "one-third-FLOP threshold" (of compute used for training the new model compared to the original model) again to distinguish between model versions and new models ([63], p. 9). Again, while any threshold embodies a degree of arbitrariness, it does provide legal clarity and seems generally justified for lack of a better alternative. Note that this threshold, and the behavioural criteria that need to be considered in addition, only apply if the potentially new model is developed by the *same* entity as the original model. A model developed by another entity will generally qualify as new.

3.3.2. Developing novel architectures

One clear case of new model creation involves the development of entirely new architectures, particularly those that depart from the current transformer-based paradigm for LLMs. For example, a developer who designs and implements a model based on Joint Embedding Predictive Architectures (JEPA) [81,82] or another non-transformer foundation [83], while the original, inspirational model is transformer-based, is not simply modifying an existing system but creating a new one. In such cases, the resulting model will likely exhibit both architectural and behavioural novelty, and the developer must show full compliance with the AI Act's obligations for providers.

3.3.3. Distillation

A more complex case arises in the context of model distillation. Distillation involves training a smaller "student" model to approximate the outputs of a larger, pre-trained "teacher" model [49–51]. While the goal is to retain similar behaviour, the distilled model may develop distinct operational properties–such as differences in "reasoning" reliability, "hallucination" frequency, or sensitivity to inputs–due to its reduced size and changed training dynamics. This divergence can lead to changes in the risk profile and alignment properties of the student model.

As a result, distillation generally results in what should be treated as a new model under the AI Act, thus triggering provider status and associated compliance duties for the entity performing the distillation. This holds *a fortiori* if the student model is distilled by a company that is different from the entity that develops the teacher model. For example, if DeepSeek indeed distilled some of its models from OpenAI models (cf [69,84,85]), they clearly do constitute new models, even if their behaviour closely matches those of the teacher models.

The development of an entirely new AI model triggers full provider obligations under the EU AI Act, but does not invoke collaborative duties under Article 25(2). A model qualifies as new if it introduces substantial architectural changes, displays qualitatively different behaviour, or is trained from scratch by a different entity, including through techniques like distillation. Even when behavioural differences are minimal, a new developer or significant training efforts–such as passing the one-third FLOP threshold proposed by the AI Office–generally supports the classification as a new model.

4. Modifications under AI liability

The AI Act is not the only framework in AI regulation. Both in the EU and globally, an equally important but often overlooked component is AI liability, which is not covered by the AI Act [67,86–89]. Hence, this section examines how liability attaches to modifications of AI systems under the recently updated EU framework. Although the analysis focuses on EU law, product liability and tort law are globally relevant and form a legal background widely shared in many countries. Even in jurisdictions like the United States, AI remains subject to general tort law and product liability, despite the absence of dedicated AI legislation at the federal level [90,91].

Under EU law, both the recently revised Product Liability Directive (PLD) and Member State tort law govern liability for GPAI systems. Concerning the Member State regime, the European legislators had contemplated the introduction of an AI Liability Directive [86,87,92,93]. However, this proposal was, despite significant support in the European Parliament for AI liability reform [94], withdrawn by the European Commission [95], potentially in a hasty and ill-conceived attempt to "simplify" AI regulation. While the withdrawal leaves a significant gap in liability coverage and particularly enforcement, the provider of a GPAI model may still be subject to liability either under the revised PLD (and the Member State transpositions) for defective products or under general tort law for breach of a duty of care. This includes liability for failure to implement safeguards against foreseeable product alteration and misuse [96–99],[6] warn of widespread modification [100], and even guard against foreseeable misbehaviour by users [101,102].[7]

4.1. Modifications under the product liability directive

The PLD was revised in late 2024 and the amendments, including its explicit application to software including artificial intelligence systems and models, need to be transposed by Member States until December 2026. The update contains new rules and defectiveness, and evidence disclosure mechanism in favour of injured parties, and presumptions for defectiveness and causality in specific situations, including the use of

[6] See for example: Welch Sand & Gravel, Inc. v. O & K Trojan, Inc., 668 N. E.2d 529, 533 (Ohio 1995) ("While a manufacturer is not responsible for all product misuses, failure to design a product to prevent a foreseeable misuse can be a design defect.").

[7] See, e.g., from Germany, BGH, Judgment of February 21, 1978 – Case VI ZR 202/76; for warning duties under product liability law against foreseeable misuse, see BGH, Judgment of May 18, 1999 - VI ZR 192–98.

machinery (Recital 48). The PLD's scope, however, is limited. It only applies in cases of specific damage categories: death and personal injury, damage to consumer property, and destruction or corruption of consumer data (Art. 6 PLD). Hence, if an AI-driven vehicle or drone hurts or kills a person, crashes into a private home, or an image modification software deletes a consumers' images, the PLD may provide redress. Conversely, many cases involving harm from generative AI – from non-discrimination to personality rights and professional intellectual property – are not covered by the PLD [89,92].

Liability becomes more complex when a downstream actor modifies the model [103]. Under the PLD, such a customizer bears liability only if they qualify as a "manufacturer." Article 8(2) PLD now specifies that a natural or legal person who substantially modifies a product outside the control of the original manufacturer and subsequently makes it available on the market shall be considered a (new) manufacturer.

Article 4(18) PLD elaborates the concept of "substantial modification." The article defines a substantial modification in two alternative ways, depending on the existence of relevant Union or national product safety rules. Under point (a), a modification is substantial if it meets the threshold established by applicable product safety legislation. Recital 39 connects this assessment explicitly to the General Product Safety Regulation (EU 2023/988, GPSR). Recital 40 extends these principles to software updates and the continuous learning of AI systems. In Article 13(3), the GPRS in turn deems a change substantial if it was not foreseen in the original risk assessment; alters the nature of the hazard or increases the level of risk; and was not carried out by the end user. Hence, this approach ties a substantial modification under the PLD directly to the broader regulatory framework governing product safety, which is harmonized across Member States through the new GPSR.

It raises, however, the tricky question of the relationship between the GPSR and the AI Act. The GPSR covers consumer products, more specifically any item intended for consumers (Article 3(1) GPSR). However, unlike in the PLD product definition, the GPRS definition does not explicitly qualify software and AI as a product; hence, one could argue that only software and AI embedded in a consumer product is covered by the GPSR, not standalone software or AI (cf. also Art. 15(3)(a) and Recital 25 GPSR) [104]. The European Commission, however, has issued guidance stating that both tangible and intangible products, including software and apps, fall under the GPSR [105]. The AI Act, in turn, covers all types of AI, constitutes specific product safety legislation [106,107] and must, hence, be considered *lex specialis* to the GPSR insofar as their scope overlaps and as the AI Act does contain specific rules (see also [63] a, para. 144; [107], Table 1). This means that, instead of Article 13(3) GPSR, the definition of substantial modification in Art. 3(23) AI Act is dispositive in the context of Article 4(18)(a) PLD. That solution affords the additional advantage of fully aligning the concepts of substantial modification in the AI Act and the PLD, which is highly desirable from a coherent and simplification perspective. It does not make sense to use one concept of substantial modification when the AI is embedded in a consumer product, and another one in all other cases. Hence, whenever the modification qualifies a substantial under the AI Act, and triggers duties for a new provider, that provider automatically is considered a manufacturer under the PLD (if the other elements of Art. 8(2) PLD are fulfilled).

Further provisions in the PLD complete the framework for such new manufacturers. Article 11(1)(g) PLD limits liability to the part of the product affected by the modification. Hence, the new manufacturer is not liable if the defectiveness does not originate in the altered part. Conversely, Article 11(2) PLD denies a liability exemption where defectiveness stems from a substantial modification *within* the manufacturer's control even if the modification occurred after the product was first placed on the market. Article 17(1)(b) PLD, finally, resets the expiry period when a substantial modification occurs.

Thus, EU product liability law recognizes a concept of "substantial modification," which implicitly references the AI Act's notion and aligns the status of new manufacturer with that of a of a new provider. Hence, an entity deemed a new provider under the AI Act will qualify as a new manufacturer under the PLD – very welcome and coherent solution.

4.2. Modifications under tort law

The PLD regime, however, only covers specific types of harm, as mentioned: death, personal injury, damage to consumer property, and loss of consumer data. Other harms–including discrimination, defamation, libel, or other violations of personality rights–fall under Member State tort law. Similarly, actors along the value chain that do not qualify as producers or manufacturers under the PLD are covered by national tort law; and even entities to which the PLD applies are subject to national tort law in addition (Art. 2(4)(b) PLD). These rules, however, are only very partially harmonized, largely limited to the area of non-discrimination and certain aspects of intellectual property law. In all other domains, national tort law typically relies on duties of care, such as the general duty not to harm others. Any customizer of an AI model must, therefore, adopt reasonable safeguards and monitoring procedures to comply with the obligation not to cause harm to others when deploying potentially dangerous products – irrespective of any duties under the PLD or the AI Act. In the context of non-discrimination law, liability is even strict, which means that any significant and causal contribution to discrimination leads to (joint and several) liability [24], irrespective of a breach of duty or fault.

In the context of general tort law, two specific interactions with the AI Act can arise. First, one may expect that courts will look to the AI Act, and particularly the GPAI and high-risk rules, to define duties of care – arguably, even beyond activities qualified as high-risk under the AI Act. This is because these GPAI and high-risk rules are largely inspired by industry best practices, which in turn typically influence duties of care under general tort law (see, e.g., [108], para. 556). Hence, to the extent that modifications lead to entities qualifying as providers under the AI Act, the respective provider duties triggered under the AI Act may indeed resurface in civil law tort claims.

Second, several national tort laws allow for civil damage claims in case specific statutory laws protecting individual interests were breached (e.g., § 823(2) German Civil Code (BGB); Art. 6:162 – Civil Code (Burgerlijk Wetboek); cf. also § 1311 Austrian Civil Code (ABGB)). The CJEU has clarified that in the case of EU regulations this is the case if the provisions, in addition to public interests, also protect the specific interests of the individual affected persons.[8] The AI Act clearly qualifies as such a protective norm. According to its Recital 1, it aims at "ensuring a high level of protection of health, safety, fundamental rights" and aspires "to protect against the harmful effects of AI systems in the Union," inter alia. Again, this means that a breach of the AI Act can lead to immediate follow-on litigation under national tort law.

Even beyond these direct influences of the AI Act, though, national tort law will find ways to hold the value chain actors accountable in terms of substantive law. A landmark German case illustrates how national courts are already addressing the liability of developers, but also deployers, for rights-violating outputs produced by automated systems. In the Autocomplete judgment, the German Federal Court of Justice (BGH) held that Google could be held liable for defamatory autocomplete suggestions generated by its search engine [109,110].[9] Google, in this context, operated both as a developer and deployer. The court ruled that although the autocomplete feature operated rather autonomously, the company bore a duty to act once it became aware of potentially harmful suggestions.[10] In one of the cases decided by the courts, the wife

[8] CJEU, Judgment of 21 March 2023, Case C-100/21, QB vs Mercedes-Benz Group AG, ECLI:EU:C:2023:229, para. 97; see also BGH, Case VI ZR 592/20, Judgment of 23 January 2024, para. 14-15.

[9] Autocomplete [2013] BGH, Case VI ZR 269/12, Judgment of May 14, 2013.

[10] Autocomplete [2013] BGH, Case VI ZR 269/12, Judgment of May 14, 2013, para. 39.

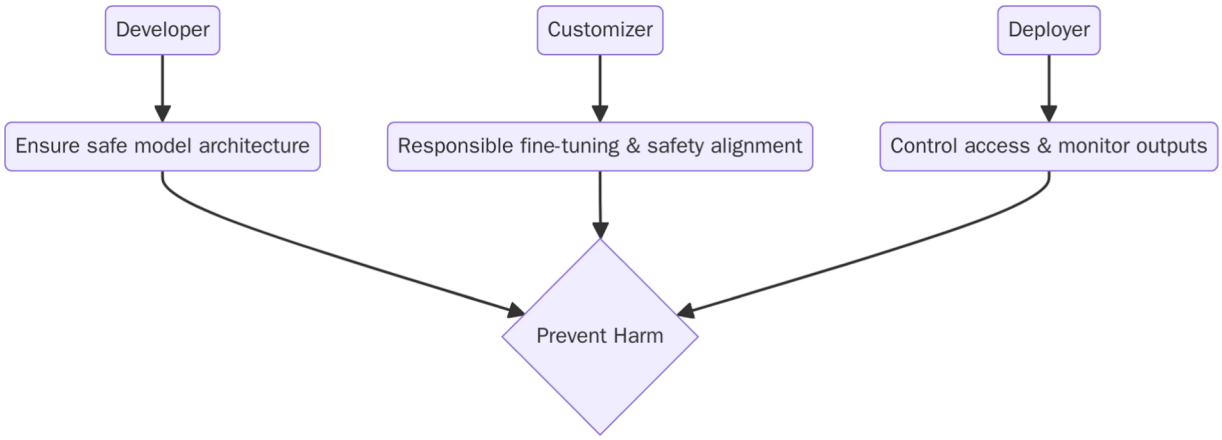

Fig. 3. Tort responsibilities along the AI value chain.

of the then-President of Germany, Christian Wulff, sued Google after the term "prostitute" appeared as a suggested search term following her name. The court emphasized that, upon notice, the provider had an obligation to undertake reasonable and appropriate steps to prevent further rights-infringing outputs, based on a duty to mitigate once harm becomes foreseeable. Structurally, this is similar to the notice-and-takedown obligations under copyright and platform law.

This precedent is particularly relevant in the context of generative AI models, especially LLMs, which function as complex and autonomous text-generation systems–essentially large-scale autocomplete engines. While Google's autocomplete relied on a much smaller model at the time (a small language model), the same underlying principle applies today, in our view: if developers fail to implement adequate safeguards or fail to act upon notification of harmful outputs, they may incur liability for the violation of personality or other rights [24]. This liability is not limited to developers. Deployers and customizers of AI systems–those who integrate, fine-tune, or otherwise modify or make generative models publicly accessible–are also bound to be held to such standards by courts.

Essentially, each party in the AI value chain needs to do whatever is possible, reasonable, and appropriate given its control over the system and its output to prevent harm (see Fig. 3). These steps include integrating moderation systems or guardrails and conducting testing during training and modification activities, proportionate to the size of the training or modification; evaluation of appropriate samples of outputs; and establishing effective notice-and-takedown procedures if models are made available by professional actors vis-à-vis end consumers. Downstream actors may further insulate themselves from liability by setting up robust compliance mechanisms, which could involve prompt response protocols for removing or blocking harmful content and logging intervention steps to show diligence.

4.3. Allocating liability along the value chain

The question of liability attribution between the original and the new manufacturer or customizing entity follows the pattern of causation. If the unmodified model does not generate the harmful output–at least not with statistical significance–then the modification must be the causal factor. In that case, only the customizer bears liability. Conversely, if the original, unmodified model produces the harmful output in a statistically significant number of cases, then liability falls on the original provider. In scenarios where both the original model and the modification contribute to the harm, both parties may be jointly and severally liable. For example, if the customizer exacerbates the risks created by the base model, then both parties may have breached their respective duties of care, and their actions may both qualify as causally relevant (Cf [111]).

We should note, however, that the burden of proof varies between the different legal bases. While the PLD has introduced a far-reaching reversal of the burden of proof for cases involving machine learning (Article 10 PLD, particularly para. 4 in conjunction with Recital 48), the burden of proving fault typically lies with the claimant under national tort. Particularly in cases involving complex value chains, meeting that burden of proof will be onerous, and sometimes impossible, for individual claimants. Again, this is a deficiency that the non-withdrawn AI Liability Directive was supposed to address.

4.4. Irrelevance of high-risk status under the AI Act

Importantly, the tort law obligations, and the PLD obligations to prevent product defects, arise independently of the classification of a model as high-risk under that AI Act. Of course, models used in scenarios with elevated risk need to be developed and monitored with greater care to avoid liability. However, this risk-based adaption of duties of care does not depend on the rather rigid AI Act high risk categories, but on the probability and expected severity of harm in deployment, assessed flexibly at the moment of the respective activity (development, fine-tuning, monitoring, etc.). Hence, AI liability significantly expands the duties of non-high-risk providers and users, which only face transparency and literacy obligations under the AI Act.

In summary, we note that the PLD applies only to specific harms, while national tort law governs other injuries, requiring AI model customizers to implement safeguards and respond to foreseeable harm. Courts, as in the German Autocomplete case, already hold developers and deployers liable for rights violations by automated systems if they fail to act upon notice. Liability extends along the AI value chain. Each party needs to do what is feasible, reasonable and appropriate to prevent harm to third parties. The scope of these obligations, in turn, depends on the extent that each party has control over the system. Joint liability arises when both original and customized models contribute to harm. Notably, tort and PLD duties exist independently of AI Act risk classifications, focusing instead on actual risk and severity in the expected deployment context. This significantly expands duties for non-high-risk AI providers and customizers vis-à-vis the AI Act.

With respect to specific modifications under the AI Act and the PLD, we summarize our findings in Table 2 below.

5. Policy and managerial implications

The diffusion of general-purpose AI systems into diverse domains presents new challenges and responsibilities across societal, regulatory, and organizational contexts. Based on the above discussion, this section outlines key implications for policymakers, legal frameworks, and institutional governance, with a focus on balancing innovation with

Table 2
Summary of GPAI model modifications under the AI Act and the PLD.

Modification Type	Classification	Provider (AI Act) and Manufacturer Status (PLD)	Reasoning
System Prompt Changes	Generally insubstantial	No (unless major risk)	Does not alter architecture or intended purpose. May trigger provider status only if prompt introduces high-risk behaviour (e.g., hate speech).
CustomGPTs	Generally insubstantial	No (unless major risk)	Same as system prompts; customization does not change weights. Provider status only if consequence scanning reveals elevated risk.
RAG with general data sources	Generally insubstantial	No	Retrieval-based enhancement without changes to core model. Behaviour not substantially altered.
RAG with Sensitive Sources (e.g., nuclear safety data)	Possibly Substantial	Yes, if major risk detected	While technically a non-architectural modification, consequence scanning may reveal high risk due to source sensitivity.
Fine-Tuning with Minor Adjustments	Generally insubstantial	No (unless major risk)	Slight parameter changes with limited data. However, provider status if consequence scanning shows emergent misalignment or increased risk.
Fine-Tuning with Biased or Risky Data	Substantial	Yes	Even with limited compute, fine-tuning can lead to new risk profile. Qualitative change in behaviour justifies provider status.
Fine-Tuning with Substantial Changes	Substantial	Yes	Large-scale retraining with broad effect on behaviour or performance. Provider status also applies if 1/3 FLOP threshold is crossed.
RLHF / RLAIF	Substantial	Yes	Updates model's internal weights and alters value alignment. Typically exceeds behavioural threshold for substantial modification.
Jailbreaking via parameter manipulation	Substantial	Yes	Alters internal safeguards or filters, directly affecting compliance and safety. Provider status applies absent compelling risk mitigation.
Distillation	New Model	Yes	Though modelled on an existing system, student model may diverge in behaviour or performance. Treated as new model due to new risk profile and large novel training run.
Development of novel architectures	New Model	Yes	New structure and/or learning paradigm; not derived from an existing system. Full provider obligations apply.

accountability, and flexibility with enforceability.

5.1. Policy implications

AI technologies now influence core public interests, from democratic integrity to individual rights and social trust. This subsection identifies areas where regulatory clarification and policy development are urgently required, particularly in relation to provider responsibilities, transparency incentives, and legal responsibility across the AI value chain.

5.1.1. Clarification on provider status for modified GPAI models

The regulatory treatment of modified GPAI models requires clarification, particularly regarding new provider status and new models. A delegated act issued by the European Commission would constitute the most effective and authoritative mechanism to provide this clarification. Such a mandate could be included in an AI Act update, for example in the contemplated Omnibus regulations. If such an act is not feasible, then detailed guidance documents should be developed to ensure legal certainty and regulatory consistency; at the time of writing, they seem indeed to be forthcoming.

Going forward, a dual-track evaluation framework should be adopted to differentiate between varying degrees of model modification. This would allow proportionate regulatory obligations and reduce compliance burdens where appropriate. It would replace the current dichotomy between either qualifying entirely as a new provider, or not at all. We suggest the following intermediate obligations for customizers:

- Track A (Substantial Modifications): Entities that introduce substantial modifications to GPAI models must conduct proportionate ex-ante risk evaluations. These evaluations must assess whether the changes significantly alter the model's risk profile in ways that could affect public safety, fundamental rights, or systemic stability.
- Track B (Minor Modifications): Minor technical or functional changes may proceed without pre-market model testing. Instead, these cases should require internal consequence analysis (essentially covered by the CCS test), documentation, and adherence to applicable tort law norms. Civil liability principles must be adequately enforceable to provide redress in the event of harm, which speaks in favour of resurrecting the AI Liability Directive (Hacker, 2024; [67]).

To determine whether a modification qualifies as substantial or minor, the CCS test developed above can be used. The novelty vis-á-vis the current rules then lies in the legal consequence, which would lie between the full GPAI obligations currently facing new model providers as one extreme, and the lack of any model-related obligations for insubstantial modifications as the other.

5.1.2. Voluntary model reference repository

A central pillar of federated compliance is the capacity to reference the base FM when assessing the origin of a system's behaviour. This principle underlines the importance of establishing a reference database to house original, unmodified versions of GPAI models. While current

regulatory frameworks–such as Article 71 of the EU AI Act–only mandate registries for *high-risk* AI systems and include merely metadata, there is a strong case for extending this framework. Specifically, we propose the creation of a voluntary reference repository for GPAI providers to submit reference versions of their models. Technically, this is feasible: although uploading training datasets remains impractical due to size and sensitivity, uploading model artefacts–ranging from gigabytes to terabytes depending on precision (e.g., FP32, FP16, INT8)–is well within reach. Such a reference repository would simplify the attribution of responsibility in the event of harmful or unlawful AI behaviour by enabling a straightforward benchmark: does the original, unmodified model exhibit the same fault?

From a compliance standpoint, the repository would serve as an evidentiary safeguard for providers. In the absence of a registered, original reference model, a civil liability presumption (concerning product defect under the PLD and fault under national tort law) could be made against the original FM provider–placing the burden on the FM provider to demonstrate that the base model does not produce the incriminating output. We would imagine such a presumption to require a novel statutory basis: in EU law for the PLD, and in a harmonizing EU act or in national laws for national tort laws. In the absence of such as statutory presumption, it seems unlikely that courts would derive such presumption from the mere existence of a voluntary repository. Importantly, such a presumption would not constitute a de facto mandate to disclose the full model. Providers could rebut liability by offering sufficient verifiable documentation, or by presenting evidence to a neutral third party under non-disclosure arrangements, which would preserve trade secrets and intellectual property.

Extending this model, a complementary voluntary registry for entities that modify GPAI systems should also be encouraged. This secondary registry would record the core features of modified models, enhancing transparency, traceability, and accountability. Participation would signal a commitment to responsible AI practices, promote peer learning, and enable the sharing of safety-relevant data across stakeholders. Hopefully, these registries would foster a culture of openness while acknowledging the need to protect some proprietary technologies–ultimately advancing a more trustworthy and governable AI ecosystem.

5.1.3. Towards a legal regime for AI intermediaries

More generally, the emergence of AI intermediaries–such as model customizers, hosting providers, and application-layer developers–demands the articulation of a coherent legal framework that reflects their distinct roles in the AI value chain and infrastructure stack. A structured approach must recognize the triangle of legal regimes requiring revision or expansion: AI-specific obligations, platform responsibilities, and criminal liability in cases of severe abuse.

First, AI-specific duties must apply to entities that customize or fine-tune GPAI models. As discussed above, these actors must be treated as specific entities under the AI Act with specific duties, different from but similar to those of providers, when modifications materially influence the model's risk profile. This classification entails compliance with relevant pre-market assessments, documentation, and post-market monitoring obligations.

Second, intermediary liability and platform governance duties, similar to those established under the Digital Services Act (DSA), should be extended to AI hosting services and application providers [112]. These entities must ensure a reasonable degree of safe deployment, use, and accessibility of AI systems. Duties may include certain content moderation obligations, transparency reporting, and mechanisms for flagging and mitigating harm, especially where outputs are widely distributed or user-generated (see, e.g., [4,5]).

Third, criminal liability must be updated and enforceable against all actors–regardless of their position in the supply chain–when their systems are used to commit serious offenses and they have violated duties considered fundamental enough to be sanctioned by criminal law. This includes, but is not limited to, the production or distribution of non-consensual intimate imagery (NCII) [113], child sexual abuse material generated through image models [30,114,115], or fraud committed via AI-generated content [116]. In this context, national criminal laws must be updated to ensure they apply effectively in AI-mediated environments [117].

A fuller articulation of these three complementary legal layers would close critical regulatory gaps across what Michael Veale and Robert Gorwa call the regulation of the "AI Stack",[11] distribute responsibility appropriately across the ecosystem, and provide clear avenues both for compliance and for redress in cases of harm.

5.2. Managerial implications

A range of practical challenges remain regarding how organizations should assess their legal obligations, particularly in relation to the extent of modifications made to foundational AI models, and subsequently how compliance protocols and procedures should be structured and implemented. Technical standards currently being developed by JTC21 of CEN—CENELEC offer a potential pathway, as Articles 40(1) and 55(4) of the AI Act provide organizations the possibility of claiming a "presumption of conformity" by adhering to these standards. Nonetheless, there are significant limitations associated with these emerging standards, as highlighted by Kilian et al. [118]. They are substantially delayed and not expected to be finalized until mid- to end of 2026 at the earliest, exhibit an overrepresentation of large technology firms that may limit their applicability to small and medium-sized enterprises (SMEs) and sectors not specializing in AI, and often lack sufficient specificity. Given the absence of robust and applicable standards, we propose the following actionable approach.

5.2.1. Structured risk assessment using FMEA and consequence scanning

Organizations should consistently apply Failure Mode and Effects Analysis (FMEA) as a structured methodology to evaluate potential risks associated with AI systems. Consequence scanning, which systematically examines impacts on all stakeholders, provides an effective approach to operationalizing FMEA. Such comprehensive stakeholder-focused evaluation is essential for thorough risk assessment and mitigation.

5.2.2. Implementation of state-of-the-art protocols

Organizations should adopt best practices and adhere to state-of-the-art methods relevant at the time of testing and release. Compliance with Article 8 of the AI Act, which mandates adherence to "generally acknowledged state-of-the-art" methods for high-risk AI systems, is strongly supported by such practices. Recital 64 further emphasizes this requirement. Article 9(3) suggests that risk mitigation must align with current best practices and generally recognized technological standards (see also Recital 65). Additionally, Article 50(2) specifies labelling requirements for synthetic content, explicitly referring to technological and market developments (Recital 133). Furthermore, evaluations under Article 55(1)(a) stipulate that standardized, state-of-the-art testing protocols be used. While SMEs enjoy reduced obligations, universally, aligning with best practices substantially supports overall regulatory compliance.

Such adherence provides three significant legal advantages. Under Article 55(4) of the AI Act, adherence may confer a "presumption of conformity," simplifying regulatory compliance. In the context of tort law, following established codes may demonstrate "reasonable care." In product liability, it indicates that no superior, reasonable alternative design was available. This makes a breach of duty (tort) and a product

[11] See, e.g., https://quello.msu.edu/event/full-stack-ai-governance-with-robert-gorwa-wzb-berlin-social-science-center-and-michael-veale-university-college-london/.

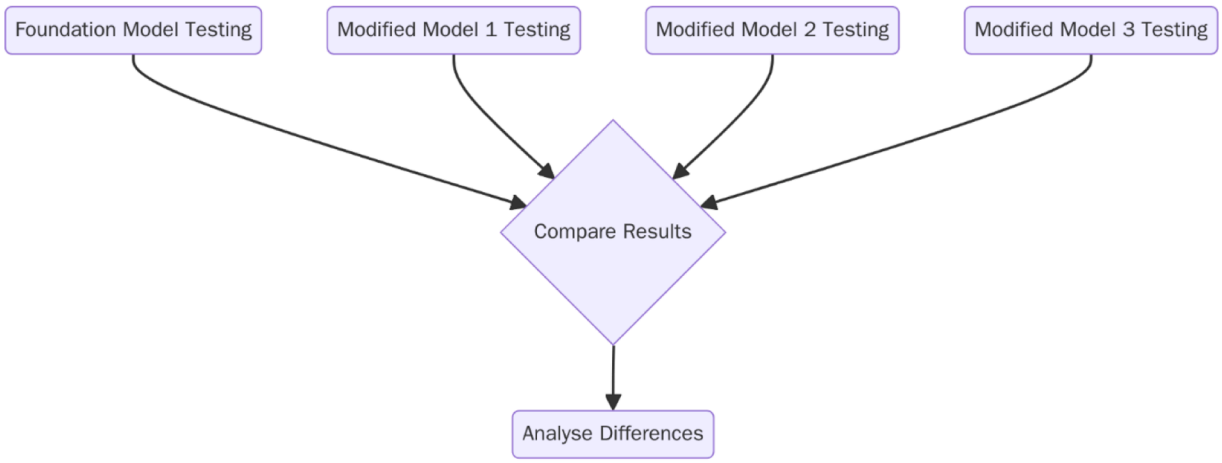

Fig. 4. A scalable federated compliance regime.

design defect (product liability) unlikely.

5.2.3. Comprehensive documentation and audit trails

Organizations should maintain thorough documentation to mitigate risks associated with the reversed burden of proof concerning defects in machine learning models and systems under the Product Liability Directive [87,89]. Creating and managing a comprehensive AI inventory and maintaining an audit trail through governance platforms ensures that evidence of responsible compliance practices is preserved and may protect organizations in potential liability cases.

5.2.4. Federated compliance and comparative testing

As stressed throughout this paper, compliance for modified general-purpose AI systems benefits from a federated approach involving multiple actors across the value chain. This comprises conducting systematic comparative testing: initially testing the original foundation model, repeating the same tests on the modified model, and then analysing any differences identified. Effective federated compliance thus relies on active collaboration and information sharing with the original GPAI model providers, or at least on access to the original, unmodified model. Notably, this approach scales efficiently, maintaining a consistent testing architecture regardless of the number of participants involved (Fig. 4).

5.2.5. Examples

Consider the fourth scenario mentioned in the introduction, in which a chatbot is derived from a foundation model to assess car insurance eligibility. If the discriminatory outcomes are present in the original, unmodified model, then the provider of that original model is ultimately responsible. If, however, the discriminatory output only surfaces in the modified model(s), the problem must be in the fine-tuning data set, and responsibility lies with the modifying entity. This is of crucial importance for the internal redress between modifier and original provider.

Typically, the test also applies with respect to liability and obligations vis-à-vis third parties. For example, in the first, second, and third scenario contemplated in the introduction ("hallucinations" due to fine-tuning; harmful advice due to misaligned RLHF; harm to workers due to outdated RAG database), the original provider of the foundation model did not violate the AI Act (with respect to the specific problem mentioned) and is not liable to the injured parties under the PLD or tort law as the original provider did not breach any duty of care or put a defective model onto the market.[12]

[12] This assumes that the original provider does not have any control over the modified model anymore ad instructed the users adequately.

The discrimination case is more difficult as liability is strict in non-discrimination law. However, non-discrimination law does not provide any explicit guidance on who should be liable in an AI value chain—which does not come as a surprise since the European directives date to the 2000′s, when algorithmic discrimination was not high on the legislators' minds. As one of the authors of this paper has argued, the structure and purpose of non-discrimination law likely dictates that, typically, both the provider and the deployer of an unmodified model are liable if that unmodified model delivers discriminatory output [anonymized]. However, in our view, the chain of attribution is broken if the original model is substantially modified. Ultimately, it cannot be the case that all entities who causally contributed to the discriminatory modified model, up to the electricity provider, are liable under non-discrimination law. Hence, it makes sense to demand that, in case of model modifications, the provider of the original model is only liable if the original model produces the same or highly similar discriminatory output as the modified one. In these cases, the original provider is jointly and severally liable with the customizing entity and the deployer, just like the original provider with the deployer in the case of unmodified models (external liability allocation). However, deployer and customizing entity can fully claw back from the original provider any damages they may have had to pay to the injured party (internal liability allocation).

5.2.6. Reduced mandates for low-risk GPAI and SME providers

We acknowledge that the comprehensive compliance approach outlined above demands significant technical expertise and resources, potentially limiting its accessibility primarily to larger organizations. This extensive requirement could indeed be prohibitive for smaller entities. However, SMEs that primarily customize off-the-shelf AI models can adopt a streamlined version of this approach by focusing on consequence scanning, adherence to established codes of conduct, and documentation via templates. While SMEs benefit from reduced regulatory obligations under the AI Act, it is critical to emphasize that their liability under product liability and tort laws remains generally unaffected by organizational size. Consequently, and rightly, SMEs must still diligently manage liability risks associated with deploying unmodified and customized AI systems.

6. Conclusion

Our analysis highlights two critical findings that shape the approach to compliance and liability concerning general-purpose AI systems.

First, effective compliance management for GPAI must adopt a federated approach. Due to the inherently distributed nature of GPAI development and deployment, responsibility for compliance should not

be centralized but rather shared across multiple actors within the value chain. Importantly, the specific regulatory responsibilities and mandates for each actor depend significantly on the extent and nature of the customization applied to the foundational model. To determine whether an organization becomes a provider under the AI Act, we suggest the CCS test, which combines the bright red line of a Compute threshold with the moderate burden of functional risk analysis via Consequence Scanning.

Second, we emphasize that compliance with AI-specific regulations, such as the AI Act, should not be considered in isolation from broader product liability considerations. Regulatory compliance and liability risk are deeply interconnected. While regulatory frameworks primarily focus on the nature and extent of customization to define compliance responsibilities, liability frameworks emphasize the actual consequences and risks posed to stakeholders. In practical terms, organizations must concurrently assess their regulatory status (such as provider obligations under the AI Act) alongside their broader liability exposure resulting from the real-world impacts of deployed AI systems.

We thus argue that organizations should leverage tools such as voluntary codes of practice and rigorous state-of-the-art testing, as these can provide crucial legal protections and significantly mitigate liability risks. Recognizing the intertwined nature of regulatory and liability frameworks is essential to effectively managing the complexities inherent in deploying customized general-purpose AI systems.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical approval and informed consent statements

Not applicable. The study did not involve human participants or animals that would require ethical approval or informed consent.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

None.

Data availability

No data was used for the research described in the article.

References

[1] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., & Brunskill, E. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
[2] Fernández-Llorca D, Gómez E, Sánchez I, Mazzini G. An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. Artif Intell Law 2024;33(4):1–14.
[3] Gstrein OJ, Haleem N, Zwitter A. General-purpose AI regulation and the European Union AI Act. Internet Policy Rev 2024;13(3):1–26.
[4] Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: Proceedings of the *ACM conference on fairness, accountability, and transparency (FAccT '23)*; 2023. p. 1112–23. https://arxiv.org/abs/2302.02337.
[5] Helberger N, Diakopoulos N. ChatGPT and the AI Act. Internet Policy Rev 2023;12(1):1–6.
[6] Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI Ethics 2024;4(4):1085–115.
[7] Ohm P. Focusing on fine-tuning: understanding the four pathways for shaping generative AI. Sci Technol Law Rev 2024;25(2):214–40.
[8] Schuett J. Risk management in the artificial intelligence act. Eur J Risk Regul 2024;15(2):367–85.
[9] Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Ó hÉigeartaigh, S., & Korinek, A. (2023). Open-sourcing highly capable foundation models: an evaluation of risks, benefits, and alternative methods for pursuing open-source objectives.
[10] Williams, S., Schuett, J., & Anderljung, M. (2025). On regulating downstream AI developers. *arXiv preprint arXiv:2503.11922*.
[11] Anisuzzaman D, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digital Health 2025;3(1):100184.
[12] Blaschke T, Bajorath J. Fine-tuning of a generative neural network for designing multi-target compounds. J Comput Aided Mol Des 2022;36(5):363–71. https://doi.org/10.1007/s10822-021-00392-8.
[13] Laufer B, Kleinberg J, Heidari H. Fine-tuning games: bargaining and adaptation for general-purpose models. In: Proceedings of the ACM web conference 2024; 2024.
[14] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
[15] Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025). Emergent misalignment: narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.
[16] Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.
[17] AI Forensics & AlgorithmWatch. (2023). Generative AI and elections: are chatbots a reliable source of information for voters? Algorithmwatch.Org.
[18] Berz A, Engel A, Hacker P. Generative KI, datenschutz, hassrede und desinformation – zur regulierung von KI-Meinungen. Zeitschrift für Urheber- und Medienrecht; 2023. p. 586.
[19] Binns R, Edwards L. Reputation management in the ChatGPT era. P. H. a. others (Ed.). Oxford handbook of the foundations and regulation of generative AI. Oxford University Press; 2025.
[20] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625–30.
[21] Ganguli D, Hernandez D, Lovitt L, Askell A, Bai Y, Chen A, Conerly T, Dassarma N, Drain D, Elhage N. Predictability and surprise in large generative models. In: Proceedings of the *ACM conference on fairness, accountability, and transparency*; 2022. p. 1747–64.
[22] Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.
[23] Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access 2023;11:80218–45.
[24] Hacker P, Engel A, Hammer S, Mittelstadt B. Introduction to the foundations and regulation of generative AI. In: Hacker P, Engel A, Hammer S, Mittelstadt B, editors. The Oxford handbook of the foundations and regulation of generative AI. Oxford University Press; 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5137750.
[25] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., & Zhang, Z. (2023). AI alignment: a comprehensive survey. *arXiv preprint arXiv: 2310.19852*.
[26] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1–38.
[27] Marcus G, Southen R. Generative AI has a visual plagiarism problem. IEEE Spectrum; 2024. https://spectrum-ieee-org.cdn.ampproject.org/c/s/spectrum.ieee.org/amp/midjourney-copyright-2666872100.
[28] Oxford Analytica. Generative AI carries serious online risks. Emerald Expert Briefings; 2023.
[29] Stokel-Walker C, Noorden RV. The promise and peril of generative AI. Nature 2023;614(1):214–6.
[30] Thiel, D. (2023). Identifying and eliminating CSAM in generative ML training data and models. Technical Report, Stanford University, 2023.
[31] Gilchrist W. Modelling failure modes and effects analysis. Int J Qual Reliab Manag 1993;10(5):16–23.
[32] Liu H-C, Liu L, Liu N. Risk evaluation approaches in failure mode and effects analysis: a literature review. Expert Syst Appl 2013;40(2):828–38.
[33] Sharma KD, Srivastava S. Failure mode and effect analysis (FMEA) implementation: a literature review. J Adv Res Aeronaut Space Sci 2018;5(1):1–17.
[34] Verma, P., & Oremus, W. (2023, April 5, 2023). ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. *Washington Post*. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.
[35] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., & Henighan, T. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv: 2204.05862*.
[36] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2023). Safe rlhf: safe reinforcement learning from human feedback. *arXiv preprint arXiv: 2310.12773*.
[37] Ogunde F. Legal large language models (LLMs): legal dynamos or "fancifully packaged ChatGPT"? Discover Artif Intell 2025;5(1):21.

[38] Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., & Freire, P. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv: 2307.15217*.

[39] Khattab, O., Santhanam, K., Li, X.L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-search-predict: composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

[40] Hacker P. Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Rev 2018;55(4):1143–86.

[41] van Bekkum M, Zuiderveen Borgesius F, Heskes T. AI, insurance, discrimination and unfair differentiation: an overview and research agenda. Law Innov Technol 2025;17(1):177–204.

[42] Wachter S. Affinity profiling and discrimination by association in online behavioral advertising. Berkeley Tech LJ 2020;35:367.

[43] Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, Thomas J, Ullmann T, Becker M, Boulesteix AL, Deng D. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. Wiley Interdiscip Rev Data Min Knowl Discov 2023;13(2):e1484.

[44] Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A. and Prakash, S., 2023. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with AI feedback. arXiv preprint arXiv:2309.00267.

[45] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. and Wang, H., 2023. Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997, 2(1).

[46] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459–74.

[47] Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., & Ma, Y. (2025). SFT memorizes, RL generalizes: a comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

[48] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: low-rank adaptation of large language models. ICLR 2022;1(2):3.

[49] Gu, Y., Dong, L., Wei, F., & Huang, M. (2023). MiniLLM: knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

[50] Polino, A., Pascanu, R., & Alistarh, D. (2018). Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.

[51] Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., & Zhou, T. (2024). A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

[52] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: MIT Press; 2017.

[53] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30:1–11.

[54] Felin, T. and Holweg, M., 2024. Theory is all you need: AI, human cognition, and decision making. Human Cognition, and Decision Making (February 23, 2024).

[55] Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature 2024;631(8022): 755–9.

[56] Du W, Zhao Y, Li B, Liu G, Wang S. PPT: backdoor attacks on pre-trained models via poisoned prompt tuning. July. IJCAI; 2022. p. 680–6.

[57] Shao, Z., Liu, H., Mu, J. and Gong, N.Z., 2024. Making LLMs vulnerable to prompt injection via poisoning alignment. arXiv preprint arXiv:2410.14827.

[58] Wolf MJ, Miller K, Grodzinsky FS. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. ACM Sigcas Comput Soc 2017;47(3):54–64.

[59] Stewart, I., Horawalavithana, S., Kennedy, B., Munikoti, S. and Pazdernik, K., 2024. Surprisingly fragile: assessing and addressing prompt instability in multimodal foundation models. arXiv preprint arXiv:2408.14595.

[60] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S. and Farajtabar, M., 2024. Gsm-symbolic: understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229.

[61] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T. Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency; 2019. p. 220–9.

[62] Blum B, Rappenglück J. Fine-tuning von GPAI-Modellen nach der KI-verordnung: eine regelungslücke für zukunftstechnologie? Comput Recht 2024;40(9):626–32.

[63] European Commission. (2025). Targeted consultation in preparation of the commission guidelines to clarify the scope of the obligations of providers of general-purpose AI models in the AI act.

[64] Gutierrez CI, Aguirre A, Uuk R, Boine CC, Franklin M. A proposal for a definition of general-purpose artificial intelligence systems. Digital Soc 2023;2(3):36.

[65] Calabresi G. The cost of accidents: a legal and economic analysis. Yale University Press; 1970.

[66] Hacker, P. (2024). Comments on the final trilogue version of the AI Act. *Available at SSRN 4757603*.

[67] Wachter S. Limitations and loopholes in the EU AI Act and AI liability directives: what this means for the European Union, the United States, and beyond. Yale J Law Technol 2024;26(3):671–718.

[68] Heim, L., & Koessler, L. (2024). Training compute thresholds: features and functions in AI regulation. *arXiv preprint arXiv:2405.10799*.

[69] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., & Bi, X. (2025). DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

[70] Wendehorst C. Art. 3. In: Martini M, Wendehorst C, editors. KI-VO: verordnung über künstliche intelligenz. Beck: Kommentar; 2024.

[71] Kurita, K., Michel, P., & Neubig, G. (2020). Weight poisoning attacks on pre-trained models. *arXiv preprint arXiv:2004.06660*.

[72] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., & Zheng, Y. (2023). Prompt injection attack against LLM-integrated applications. *arXiv preprint arXiv:2306.05499*.

[73] Liu Y, Jia Y, Geng R, Jia J, Gong NZ. Formalizing and benchmarking prompt injection attacks and defenses. In: Proceedings of the 33rd USENIX security symposium (USENIX security 24); 2024.

[74] Solaiman I, Dennison C. Process for adapting language models to society (palms) with values-targeted datasets. Adv Neural Inf Process Syst 2021;34:5861–73.

[75] Lambert, N. (2025). Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*.

[76] Lambert, N. (2025). The state of post-training in 2025. *Interconnects*. https://www.interconnects.ai/p/the-state-of-post-training-2025.

[77] Wang, Z., Bi, B., Pentyala, S.K., Ramnath, K., Chaudhuri, S., Mehrotra, S., Mao, X.-B., & Asur, S. (2024). A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216*.

[78] Roose K. Can A.I. be blamed for a teen's suicide? The New York Times; 2024. October 23, https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[79] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023). Jailbreaking chatgpt via prompt engineering: an empirical study. *arXiv preprint arXiv:2305.13860*.

[80] Wei A, Haghtalab N, Steinhardt J. Jailbroken: how does LLM safety training fail? Adv Neural Inf Process Syst 2024;36:1–32.

[81] Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, LeCun Y, Ballas N. Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023.

[82] Mo S, Tong P. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. Adv Neural Inf Process Syst 2024;37: 2348–77.

[83] Huang K, Wang Y, Zhang X, K. H. a. others. Foundations of generative AI. Generative AI security: theories and practices. Springer; 2024. p. 3–30.

[84] Deng Z, Ma W, Han Q-L, Zhou W, Zhu X, Wen S, Xiang Y. Exploring DeepSeek: a survey on advances, applications, challenges and future directions. IEEE/CAA J Autom Sin 2025;12(5):872–93.

[85] Xu, Z., Wang, J., Xu, X., Yu, P., Huang, T., & Yi, J. (2025). A survey of reinforcement learning-driven knowledge distillation: techniques, challenges, and applications.

[86] De Bruyne J, Dheu O, Ducuing C. The European Commission's approach to extra-contractual liability and AI–An evaluation of the AI liability directive and the revised product liability directive. Comput Law Secur Rev 2023;51:105894.

[87] Hacker P. The European AI liability directives - critique of a half-hearted approach and lessons for the future. 51 Comput Law Secur Rev 2023;51(2023): 105871. Article.

[88] Novelli C, Casolari F, Hacker P, Spedicato G, Floridi L. Generative AI in EU law: liability, privacy, intellectual property, and cybersecurity. Comput Law Secur Rev 2024;55:106066. https://doi.org/10.1016/j.clsr.102024.106066. Article.

[89] Wagner G. Liability rules for the digital age - aiming for the Brussels effect. Eur J Tort Law 2022;13(3):191–243.

[90] Ramakrishnan, K., Smith, G., & Downey, C. (2024). US tort liability for large-scale artificial intelligence damages. RAND Research Report.

[91] Vogel M, Chertoff M, Wiley J, Kahn R. Is your use of AI violating the law? An overview of the current legal landscape. NYUJ Legis Pub Pol'y 2023;26:1029.

[92] Hacker P. Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence: complementary impact assessment. EPRS; 2024.

[93] Wendehorst C. AI liability in Europe: anticipating the EU AI liability directive. (1868-9620). Ada Lovelace Institute; 2022.

[94] Madiega T. Artificial intelligence liability directive. Briefing, European Parliamentary Research Service (EPRS); 2023.

[95] Grozdanovski L. Non-discrimination law, the GDPR, the AI act and the-now withdrawn-AI liability directive proposal offering gateways to pre-trial knowledge of algorithmic discrimination. AI Ethics 2025:5(5):1–24.

[96] Gallub MB. Limiting the manufacturer's duty for subsequent product alteration: three steps to a rational approach. Hofstra L Rev 1987;16:361.

[97] Sachs RA. Product liability reform and seller liability: a proposal for change. Baylor L Rev 2003;55:1031.

[98] Swanson G. Non-autonomous artificial intelligence programs and products liability: how new AI products challenge existing liability models and pose new financial burdens. Seattle UL Rev 2018;42:1201.

[99] Zablotsky P. The appropriate role of plaintiff misuse in products liability causes of action. 10 Touro L. Rev 1993:183–209.

[100] Stilwell T. Warning: you may possess continuing duties after the sale of your product-(An evaluation of the restatement (Third) of torts: products liability's treatment of post-sale duties). Rev Litig 2007;26:1035.

[101] Czernik CA. Die verkehrssicherungspflicht auf privaten grundstücken: haftung auch gegenüber unbefugten nutzern? Disserta Verlag; 2010.

[102] Förster C. Verkehrssicherungspflichten. Jurist Arbeitsblätter 2017:721–8.

[103] Arcila, B.B. (2025). AI liability along the value chain.

[104] Lölfing, N. (2024). Product compliance for consumer products powered by AI. https://www.twobirds.com/en/insights/2024/global/product-compliance-for-consumer-products-powered-by-ai.

[105] European Commission. (2024). Product safety legislation. https://ec.europa.eu/safety-gate/#/screen/pages/productSafetyLegislation.
[106] Veale M, Borgesius FZ. Demystifying the draft EU artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. Comput Law Rev Int 2021;22(4):97–112.
[107] Almada M, Petit N. The EU AI Act: Between the rock of product safety and the hard place of fundamental rights. Common Mark. Law Rev. 2025;62(1):85–120.
[108] Wagner G. § 823 Bgb. In: Säcker F, Rixecker R, Oetker H, Limperg B, Schubert C, editors. Münchener kommentar zum Bgb. Band 7. C.H. Beck; 2024.
[109] Hacker, P., Mittelstadt, B., Borgesius, F.Z., & Wachter, S. (2025). Generative discrimination: what happens when generative AI exhibits bias, and what can be done about it. In P. Hacker, A. Engel, S. Hammer, & B. Mittelstadt (editors), *The Oxford handbook of foundations and regulation of generative AI*.
[110] Smith ML. Search engine liability for autocomplete defamation: combating the power of suggestion. U Ill JL Tech & Pol'y 2013(2)::313–36.
[111] Green MD. The unanticipated ripples of comparative negligence: superseding cause in products liability and beyond. SCL Rev 2001;53:1103.
[112] Gorwa R, Veale M. Moderating model marketplaces: platform governance puzzles for AI intermediaries. Law Innov Technol 2024;16(2):341–91.
[113] Hawkins W, Mittelstadt B, Russell C. Deepfakes on demand: the rise of accessible non-consensual deepfake image generators: the rise of accessible non-consensual deepfake image generators. In: Proceedings of the ACM conference on fairness, accountability, and transparency; 2025.
[114] Thiel D, Stroebel M, Portnoff R. Generative ML and CSAM: implications and mitigations 2023.
[115] Trivison A. Understanding the line between art and abuse: how generative AI changes the landscape of child sexual abuse materials. Cathol Univ J Law Technol 2024;33(1):87–114.
[116] Herbosch M. Fraud by generative AI chatbots: on the thin line between deception and negligence. Comput Law Secur Rev 2024;52:105941.
[117] Sundberg TK. Federalizing NCII regulation: the take it down Act's approach to criminalization, platform liability, and threats to disseminate. Ga L Rev 2024;59:1207.
[118] Kilian R, Jäck L, Ebel D. European AI Standards-Technical Standardization and Implementation Challenges under the EU AI Act. 2025.