Position: LLM alignment data should be regulated as mass media

João Gonçalves [1]

Abstract

Most efforts to regulate and estimate the societal impacts of Large Language Models (LLMs) are aimed at model outputs. This makes regulation difficult, because outputs are stochastic and highly conditioned on diverse user prompts. This position paper draws from media and communication literature to argue that the regulatory focus has been misplaced, and that alignment datasets (e.g., supervised fine-tuning and preference pairs) should be regulated at the same level as mass media content such as newspaper articles or television advertising. Post-training alignment data has a direct influence on all user interactions with a model, representing the same one-to-many communication flow as traditional mass media. At the same time, mass media regulation has balanced for decades the need for audience protection with room for pluralist perspectives, providing a source of learning and inspiration for LLM regulation. Regulating post-training alignment data as mass media content is the most direct and actionable route for pluralism and accountability in LLM development and deployment.

1. Introduction

Auto-regressive large language models are regularly used by a large portion of the world's population. In the European Union (EU) alone, nearly one in three individuals used generative AI tools in 2025 (Eurostat, 2025). In countries like the United States, the number of people using ChatGPT is now comparable to those who listen to podcasts, or read newspapers (Newman et al., 2025). Despite this comparable reach, the editorial choices made by developers and providers in model training are substantially less scrutinized than those of broadcasters or publishers. These differences persist in spite of public outcry about the consequences of unrestricted LLM exposure on domains such as (mental) health, education, or democratic processes (Oomen et al., 2024), for which pluralism is a prerequisite.

The assumption seems to be that, due to the non-deterministic nature of the model outputs, model providers have limited responsibility in relation to the content presented to users, much like social media platforms that argue that users are accountable for the posted content, not themselves. Some positions even argue for a closer alignment of LLM regulation with social media regulation (Appel, 2025). In contrast, this paper argues that the process of training a machine learning model, particularly post-training alignment, is an editorial process that matches much more closely journalistic editorial practices than platform services, as shown in Figure 1. Namely, it involves the selection and filtering of content, known as gatekeeping in journalism, making choices of style and presentation through Direct Preference Optimization (DPO) and Reinforcement Learning with Human Feedback (RLHF), and delivering variations of this content through model deployment to a mass audience. While platforms may indirectly select the content presented to users based on algorithmic ranking, model providers actively create, select, and adjust content in post-training data. As such, **we argue that post-training alignment data should be regulated as mass media content**.

We argue that parallels with mass media are a necessary heuristic to avoid historical pitfalls on pluralism that media regulation also faced, namely the tensions between pluralism and diversity (Raeijmaekers & Maeseele, 2015). If media (and LLMs) simply aim to represent a wide range of positions in society, for instance, by mimicking the prevalence of positions within a target audience, this might include positions that are not grounded in reason and truth (e.g. anti-vaccine, climate change deniers and other forms of conspiracy theories) and, paradoxically, positions that argue that only a single position is valid, therefore countering the pluralistic objective of the model. The content of such media systems would be diverse, but not plural. Media pluralism implies not only diversity of positions, but also procedures of exchange and debate that are grounded in objectivity, impartiality and fairness. To some extent, these

[1]Department of Media and Communication, Erasmus University Rotterdam, Rotterdam, The Netherlands. Correspondence to: João Gonçalves <ferreiragoncalves@eshcc.eur.nl>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

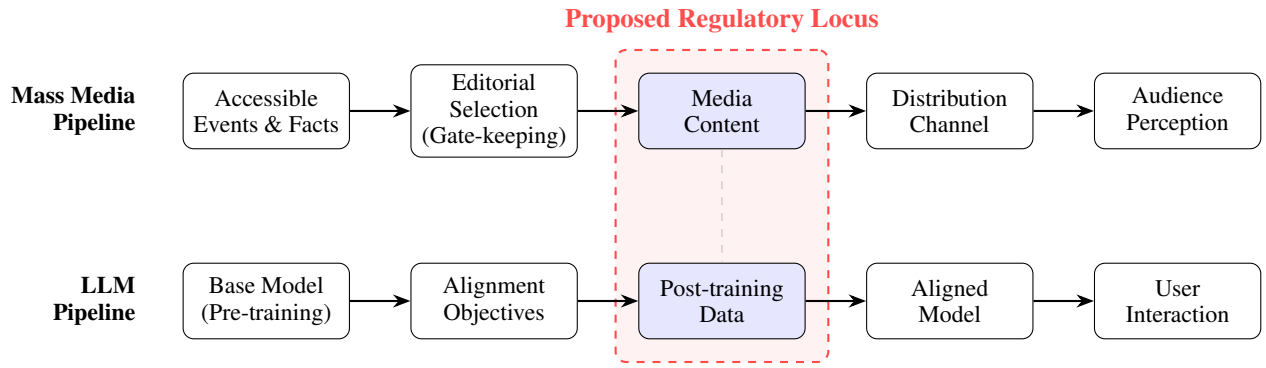

Figure 1. Comparison of Mass Media and LLM Pipelines showing the proposed regulatory locus.

could be described as the process of narrowing down all positions to a subset of *reasonable* positions (Sorensen et al., 2024). However, who determines what is reasonable and how is it determined? If we rely on model developers or benchmark creators to determine this, we may again fall under a trap where the outcomes are diverse, but not pluralistic. We therefore argue that regulation for pluralistic alignment requires broad participation of societal stakeholders, ensuring that pluralistic alignment results from public debate of a wide range of individuals and their positions. At the same time, this debate is not static and is procedural. Media norms such as objectivity, impartiality and fairness are not about how the content should be presented, but about the set of procedures that should be followed to result in pluralistic content. Therefore, regulation should not take place at the output (content) level, but at the post-training (procedure) level.

The shift in focus from model outputs to post-training data is beneficial for both societal stakeholders and model developers. From a regulatory point of view, it provides policymakers and institutions with a well established and consolidated legal framework that includes age ratings, guidelines for balanced coverage, and consumer protection. For developers, it offers concrete and consolidated guidelines for post-training data selection and (synthetic) data creation based on decades of media research and practices. There is, naturally, an argument to be made that users are not directly exposed to fine-tuning data, that pre-training might have a more determinant role in outputs (Ceron et al., 2025), and that LLM output content is still largely dependent on the input provided to the model by users. While these perspectives undoubtedly have their merits, they do not invalidate the impact that post-training has on user interactions (Weeber et al., 2026) and that, unlike unpredictable model outputs or massive pre-training data, it offers an actionable object for regulation, particularly when the goal is to ensure pluralist perspectives are represented. In response to positions stating that LLMs are a completely different technology from mass media because of their reliance on user input, we argue that communication studies have long considered the role of active audiences in traditional media through theories such as uses and gratifications (using media to fulfill specific needs) (Katz et al., 1973) or confirmation bias (people from completely opposite sides of the political spectrum might interpret the exact same text as supporting their own view) (Klayman, 1995). The role of audience views and inputs has not prevented regulation of mass media content, and neither should it prevent regulation of alignment post-training data.

In this paper we first detail the parallels between post-training datasets and mass media, focusing on related work in relation to (1) production, (2) distribution, and (3) effects, drawing on literature from the social sciences. We then outline actionable steps to transpose media regulation to LLMs by suggesting that (1) post-training data should be deposited in regulator mandated data vaults, (2) incidents should trigger data checks, and (3) regulator involvement should be proactive. We then discuss alternative viewpoints and finish with a call to action to the machine learning researchers, social scientists, and policy makers to collect, regulate, and study post-training data as media content.

2. Related work

Media studies are often divided into how content is created (production), how the content is distributed (media channels), and how it is received (media effects and reception studies). In this section we establish the parallels between stages of mass media and the role of post-training in LLM development.

2.1. Production

The first stage of mass media content production is the selection of content based on its relevance for audiences and media businesses. For journalists, this implies checking what events or information are newsworthy, either because they are relevant for the public good or because they lead audiences to purchase newspapers or visit news websites. This

process is called *gatekeeping* in media studies (Shoemaker & Vos, 2009), showing how, even before writing the first word in an article, media professionals impact audiences by selecting information. For an advertiser, this process implies selecting the product characteristics that are more appealing to the individuals in the audience, convincing them to purchase the product or service. The ways in which the selected information is presented and stylized are often hard-coded into explicit rules, such as a style guide in a newsroom or brand visual identity guidelines in advertising. These are forms of self-regulation that aim to convey clarity to both content producers and audiences, but they also often imply compliance with legal requirements such as forbidding certain kinds of sensitive content or, in the case of newsrooms, dictating the kind of language that should be used for certain politically sensitive topics (e.g., The Guardian newspaper using the wording "climate crisis" and not "climate change") (Zeldin-O'Neill, 2019), or that political parties receive proportional coverage during elections.

In a similar way, post-training alignment is often about selecting and making salient the type of interaction that is most likely to be pursued and desired by users and, for commercial models, lead to product subscriptions. Post-training is, simultaneously, the most important component in unlocking LLM capabilities and the least transparent one (Lambert et al., 2024). While newsworthiness *per se* might not be the most relevant criterion for post-training data curation, datasets tend to be tailored to the predicted uses of the model, such as coding or scientific reasoning (Fan et al., 2025). The most straightforward parallel between newsroom style guides and post-training can be found on constitutional AI (Bai et al., 2022), where harmlessness is encoded in model training through explicitly stated rules or principles instead of human labels. These editorial-like rules, however, are also enacted through post-training datasets for reinforcement learning methods such as DPO and RLHF (Winata et al., 2025) through labeler instructions and data filtering rules. In the same manner as a newsroom's style guide determines the structure of a news article, reinforcement learning approaches determine a policy that conditions the probability distribution of the model outputs. Like in media production, these principles are often expressed in textual form, such as labeling instructions for data workers or system prompts for synthetic data production. Media work is not primarily about changing the world (policy making) or the product (product design), it is about the selection and presentation of content about the world (Tewksbury & Scheufele, 2019). In a similar vein, post-training is about selecting and stylizing patterns that have been embedded in the pre-training process. However, while the former have developed (self-)regulation mechanisms to ensure that multiple perspectives are represented (e.g. journalistic source cross-checking or the fairness doctrine), the latter have no such equivalent.

2.2. Distribution

What distinguishes mass media from other forms of communication is their one-to-many delivery of content. Its origin is often traced to the Gutenberg press in the West (McLuhan et al., 2011), where technical replicability enabled the mass distribution of similar texts. Afterwards, inventions like the radio and television enabled different modalities of content, but retained the same principle of one-to-many transmission of content. While the internet and social media platforms were initially not considered mass media, being framed as many-to-many communication instead, in practice some of these platforms evolved to manifest mass media like logic (Schrape, 2016), with Instagram influencers with a large following being a notable example. Because they approximate mass media in both reach and the professionalization of content production (Van Driel & Dumitrica, 2021), these influencers are also starting to be subjected to mass media regulation such as obligations to disclosing sponsorships.

Our core argument is that post-training represents, essentially, the same one-to-many form of communication as traditional mass media. This is because instruction tuning and preference alignment are able to "constrain the model's outputs to align with the desired response characteristics or domain knowledge, providing a channel for humans to intervene with the model's behaviors" (Zhang et al., 2026). Crucially, because this is the final stage of training, the gradients of the deployed model are optimized for this specific post-training loss. For regulatory purposes, post-training data is a more comprehensive method of scrutinizing and intervening in large language models, than pre-training. An absence of pluralism in model outputs can inevitably be traced in some way or form to a pre-training corpus, but post-training alignment offers the most effective and verifiable path to address it in model development and deployment.

On the other end of the pipeline, while model outputs represent individual stochastic data points in the probability distribution of tokens, alignment data effectively shifts the distribution to approximate the dataset and associated policies. Post-training content is not delivered *ipsis verbis* to end users, and model outputs are dependent on context, but it influences model outputs to an extent that justifies framing them as one-to-many communication. This is especially true when considering that media effects, as will be seen below, depend less on explicit instances of manifest content and more on repeated exposure and latent trends. That is, even if the specific words are subject to change, effects will be strong if the underlying latent space determined by post-training is consistent.

Note that our position does not reject the claim that LLMs enable experiences that are largely personalized. The pre-

dominant chat-like format of LLM deployment means that the specific model outputs are heavily conditioned by the user prompt, and the large volume of training data and number of parameters mean that the possibilities for these outputs are vast. However, it is our position that framing LLMs as fully personalized experiences or as many-to-many communication is inaccurate, because, through alignment, model developers actively interfere at the level of the content being distributed, effectively acting as mediators of information and not simply as moderators. Regulation that ignores this active role of model developers disproportionately reduces their accountability, meaning that mass media is an adequate regulatory framework to tackle LLMs, also at the distribution level. Epistemically, large language models are not simply an aggregate of individual interactions but a force that reshapes collective intelligence through mass mediation of information (Burton et al., 2024).

2.3. Effects

The one-to-many nature of mass media led to the emergence of media effects research, aimed at questioning the consequences of media on individuals and the mechanisms through which these happen. While in the 1930s it was hypothesized that exposure to content had an immediate and predictable effect on audiences, known as the hypodermic needle theory (Ahmad et al., 2022), empirical studies have shown that media effects are often limited and conditional (Valkenburg et al., 2016), depending on factors such as repeated exposure and context. This also informs how media tend to be regulated. For instance, agenda setting theory (McCombs & Shaw, 1972) posits that issues and topics that get most attention in the media also determine what people consider to be the most important topics to them. Cultivation theory (Gerbner, 1998) states that repeated exposure to certain representations in the media lead people to adjust their perceptions of reality to match those representations, such as thinking that the world is a more dangerous place because many crime reports are shown on TV. Priming research shows that exposure to certain stimuli associated with a product can lead to subconscious changes in the reaction to that content, such as seeing an image of an attractive person before smoking or news stories about a negative topic before a campaign advertisement (Iyengar et al., 1982). The results of research on media effects are associated with regulations that aim to mitigate negative outcomes, and this is why we see pictures of unhealthy lungs on cigarette packs. Theories of agenda setting and framing have led to reporting bias and pluralism regulation (Aday, 2006), cultivation theory informs laws on sensitive content (Edwards & Berman, 1994), and priming effects are one of the justifications for bans on subliminal advertising and restrictions on labels of products like tobacco (Shi et al., 2023). While a single instance of exposure has limited effects, these regulations aim to curtail systemic and repeated exposure to harmful media content that has detrimental individual or societal effects.

In comparison to mass media, research on the effects of large language model content exposure is in its infancy (De Choudhury et al., 2023), but parallels are emerging with traditional mass media effects research. The lawsuit against Google and Character.AI related to the influence of LLMs on the suicide of a teenage boy has repeated exposure as key factor (Agence France-Presse, 2026), hinting at compounding effects from cultivation theory. In regards to media bias and pluralism, The Dutch data protection authority has issued a warning that voting advice given by chatbots for the 2025 parliamentary elections was biased and often omitted relevant political parties (Autoriteit Persoonsgegevens, 2025). These examples illustrate how problems that are traditionally considered under mass media effects research are transposed to LLM effects. Crucially, issues such as the unequal representation of political parties can be difficult to address in pre-training data, because they reflect broad and deeply ingrained societal trends, but they can be acted upon in post-training, where instruction tuning examples or preference data can steer the model to provide balanced responses for the Dutch context. However, the current focus of regulatory approaches, including the Dutch Data Protection Authority (AP) report, lies in model outputs, not post-training data. This happens because, despite the many parallels between post-training data and media content, the dominant narrative still focuses on model outputs as the key object of compliance (Sarkar, 2025), examined through resources such as benchmarks, human feedback, and red teaming. However, output benchmarks are an imperfect approximation of the post-training process, and are subject to be gamed if transparency lacks in post-training, while testing routines such as red teaming and human feedback can be vulnerable to strategic deception in model outputs i.e. *AI sleeper agents* (Sarkar, 2025). In the following sections, we outline how a shift from output regulation to post-training alignment data regulation would look like and how it can be achieved.

3. Towards a regulation of alignment data

In the previous section, we have showed the parallels between post-training alignment data and mass media content in relation to production, distribution, and effects. In this section, we show how mass media research and regulation can inform regulation of post-training data, with key ideas summarized in Table 1.

3.1. Post-training data should be deposited in regulator mandated data vaults

In traditional mass media, compliance is checked by regulators who directly examine media content. The equivalent

Media regulation	Post-training regulation	Regulatory function
Content audits	Data vault deposits	Establish accountability
Right of reply / complaints	Incident-triggered data audits	Distributes verification burden and enables pluralist scrutiny
Content guidelines	Regulator "gold" data (reference prompts / preference pairs)	Anchors compliance and enables contextualized intervention

Table 1. Translation of established mass media mechanisms into post-training regulation.

in this case would be for model providers to openly share their post-training data for inspection. This facilitates regulatory efforts, but may also imply negative effects which regulation aims to prevent. Specifically, publicly releasing post-training data related to alignment and compliance may also facilitate targeted adversarial attacks (Kumar, 2024) that circumvent the alignment purpose of post-training data. Furthermore, a significant portion of this data, such as the rejected examples in preference pairs, is harmful and highly distressing in itself, and therefore making it easily accessible to the general public may not be desirable.

Despite these objections, we argue that the one-to-many nature of post-training data requires access to it beyond the internal organization of a model provider because it carries systemic implications, similarly to mass media like the press, radio, or television required public scrutiny. Here too parallels can be established with media industries, where despite claims from industry representatives that self-regulation is efficient, accounts from practitioners show a need for additional self and state regulation (Fengler et al., 2015). The solution therefore lies in a mechanism that balances external access to post-training data and protection of intellectual property and trade secrets. We therefore propose that versioned post-training data should be deposited in data vaults with tiered access, alongside essential documentation about its provenance and use in model training. Intervening in post-training data does not automatically fix all model outputs, LLMs remain unpredictable and capable of displaying emerging capabilities beyond specific fine-tuning, but it provides a direct intervention channel to steer the direction of these models in line with legal concerns and societal priorities.

This would allow checking the data for infringing content and verifying the extent to which regulatory compliance is explicit, for instance, in preference pair data. Media regulators can then draw from the methods and standards used for traditional mass media (e.g. automated content analysis, sample auditing) to check for regulatory compliance in trusted execution environments and issue recommendations based on the outcomes of these checks. This does not necessarily imply unrestricted access to raw data. Using a text classifier for political bias on a post-training dataset would not require inspection of the actual data, only of the outputs of the classifier. For model providers, this approach has the benefit of removing some of the compliance burden, given that they can rely on scrutiny and guidance from media regulators in relation to post-training data and processes. And while objections can still be made on possible data and knowledge leaks, this approach has parallel in other industries with high public impact applications such as pharmaceuticals (clinical trial data submitted before release) and finance (a Trade Repository under the European Market Infrastructure Regulation), where regulators access highly sensitive data without jeopardizing corporate or public interest. Calls for access to non-public information from AI developers have also been made by entities like GovAI (Brundage et al., 2026), however, these typically concern safety and security risks like hacking attacks and bioterrorism, not pluralism and alignment. The access we propose does not stem from concerns such as national security, but from democratic and societal risks that are associated with a lack of pluralism. Grounded on their tradition of content and procedural scrutiny, media regulators are currently the institutions better equipped to address these risks. What enables the applicability of media regulation mechanisms is that regulatory violations would not necessarily be assessed at the output level, but at the alignment data level. The question shifts from "did model x outputs break rule y" to "did model provider x take reasonable action to prevent model x from breaking rule y".

Importantly, there is no requirement that a model provider commits to a single data vault. Differences in jurisdiction, or values, may be addressed through location-specific data vaults: where different versions of post-training data are deposited, meaning also that different fine-tunes of the model are offered based on location. This is in practice how media content is often offered, where restrictions are imposed based on the IP location because of national regulations, licensing, and cultural factors. This enables pluralism in model deployment and addresses concerns regarding the possibility that regulation by a single central authority may itself be a threat to pluralism (e.g. by globally enforcing the norms and values of a specific geopolitical context).

3.2. Users should trigger regulatory data audits

Much like in mass media content, an inherent challenge for regulation of post-training data is that the volume is too great to be manually verified by a media regulator with lim-

ited resources. Additionally, if only a subset of individuals (e.g. developers) is tasked with ensuring pluralism, this would reproduce the diversity-without-pluralism problem outlined above. Acknowledging this, media regulation has often put in place procedures and entities that members of the public can contact for complaints or issues related to media content. This includes self-regulation mechanisms such as the appointment of an independent ombudsman by news organizations, or legal mechanisms such as the right of reply when misleading information about an individual or organization is published. In effect, regulation is not only implemented by the regulator, but by the many that are targeted by media content, leveraging LLM users as a collective intelligence for participatory oversight (Burton et al., 2024).

While the public might not have direct access to the training data, they have indirect means of signaling post-training issues through their interactions with the model. Many model providers already have human feedback and red teaming efforts as part of their development process (Sarkar, 2025). If we follow the same logic as mass media regulation, this would effectively result in a systematic and centralized adversarial auditing exercise that is crowd-sourced to LLM users. If a formal complaint is made to the regulator based on outputs, they can make a targeted audit of the post-training data associated with the model to verify if there is a ground for corrective action, such as a deliberation that addresses the complaint in future versions of post-training not only for a specific model provider but for all post-training data used for models with similar purposes. As an example, if a political party files a complaint with a regulator on unfair representation of that party in model outputs during elections, regulators examine the presence or absence of that party in the post-training data through mechanisms such as keyword search or text classifiers in the data vault and, if applicable, issue a deliberation determining how parties should have a balanced representation in post-training data. Given that post-training data does not have a one-on-one representation on outputs, these audits should also examine contextual and latent concepts (e.g., not simply searching for the party names, but also for instances involving ”vote”, ”elections” or ”advice”), in the same way that a complaint on traditional mass media does not only trigger a specific correction, but also contribute to addressing systemic concerns in media system. Critically, to ensure accountability and actionable regulation, these user triggers should be always understood as a mechanism to initiate post-training data checks, never as a replacement for these checks.

3.3. Regulators should be proactive

Media regulators do not only react to incidents, but also actively issue recommendations in relation to emerging issues or high profile events. For instance, they issue guidelines about how youth should be represented in the media so they are protected from abuse (UNICEF), or what aspects should be considered to ensure balance in a televised debate. These often derive from research and monitoring conducted by the regulator, often in partnership with knowledge institutions like universities and media organizations themselves.

If post-training data is the focus of regulation, a similar proactive approach should be followed for LLMs. For instance, based on public debate and discussion, regulators could release reference contextual instructions to be included in synthetic data generation related to elections or national politics, or golden preference pairs that refine the language used when addressing youth in LLMs interactions. While a US based model provider might not have enough knowledge or resources of the political system in the Netherlands to create adequate post-training data, they could rely on the guidelines and example datasets provided by regulators to boost compliance of their models. In this way, LLM regulation, much like media regulation, ceases to be only reactive and punitive, and becomes proactive and constructive. And while regulators would have the formal authority to enforce use of these pairs, other organizations, such as NGOs, could also contribute with alignment data to the public domain in the same way, thus proactively promoting pluralism in alignment practices.

4. Societal impacts of post-training regulation

For developers and model providers, the implementation of these proposals should not bring substantial additional technical responsibilities. Limiting load for developers stemming from regulation is important to ensure that smaller organizations in the LLM space and open-source projects are not excluded from contributing to the field due to high administrative or legal burdens. The key technical requirement is that versioned alignment datasets should be deposited in location specific regulator data vaults for compliance checks and inspection. Assuming that providers already have these datasets structured for model training, depositing them on a data vault should not be substantially more complex than uploading them to public repositories such as HuggingFace. While in practice post-training pipelines include annotator guidelines, synthetic data prompts, reward-model data, safety data, preference pairs, and versioned mixtures, adding a layer of complexity to implementation, a minimal commitment to data depositing is sufficient to start a regulatory shift. This is because despite depositing the dataset being a straightforward process in itself, its main effects are not technical but editorial. Knowing that datasets will be scrutinized externally will encourage considerations about compliance during dataset curation and creation, introducing a *de facto* incentive for deliberate choices among developers and model providers. While this does require additional

investments in terms of resources, the proactive side of post-training regulation means that developers also have access to data commons contributed by regulators, an effective gold standard in relation to key compliance categories and emerging issues.

For the general public, post-training regulation offers a channel to empower LLM users to question and act on issues that they encounter during LLM use. The focus on post-training enables broader societal debates on what kind of content should be used as an example for LLM alignment and how much these examples should influence the weights of deployed models, much like the role and content of mass media are subject of political discussions and public debate. Treating post-training data as editorial content dissipates many of the black-box and accountability arguments that have prevented open discussions on LLMs. Having concrete examples of alignment data or labeling guidelines would allow for mechanisms that are already present in other forms of media to emerge for LLMs, such as community notes or letters to the editor. Taken a step further, opening up the post-training ecosystem would enable bottom-up contributions that would contribute to pluralism in a more fundamental way than top-down alignment. Examples such as citizen journalism (Wall, 2015), where individuals report on the topics or locations that are relevant to them, could also extend to alignment if developers and deployers show openness to integrate some of these contributions in their post-training pipelines.

Beyond individual contributions, post-training examples, even if not directly accessible to the public, are easily understandable and can be acted on through natural language editing and curation making them an object of discussion relevant to non-expert audiences. In the same way most people can judge and criticize an opinion piece in a newspaper, they can judge and criticize a prompt-response pair in a post-training dataset. Accountability for model behavior becomes tied to who created or synthesized the data, and to those who decided to use it for training, effectively identifying the individuals or organizations who should be involved in these public debates.

For regulators, post-training data provides a concrete locus of action to enforce compliance with newly drafted or adapted legislation and deliberations. More importantly, it enables this in a framework comparable to the regulation of television, radio, and press content, allowing media regulators to draw from their extensive track record to inform LLM regulation that promotes pluralist perspectives. On the other hand, to make effective use of the post-training data vaults, regulators should have basic knowledge and expertise on how post-training data is formatted, how it can be checked (e.g. through retrieval and classification algorithms), and its role in the LLM training process. If this is achieved, regulators effectively act as an intermediary party that protects and represents the public, while supporting model providers in their efforts for compliance.

5. Alternative views

5.1. Depositing post-training data carries unacceptable risks

In parallel with compute resources, post-training data curation and labeling is one of the most critical avenues for large language model performance (Lai et al., 2025), with top model developers investing a substantial amount of resources in protecting it. The requirement to deposit this data in tiered access data vaults may compromise the competitive advantage that market leading companies have. If regulators are compromised, or if compliance checks are abused, key characteristics of post-training recipes might leak to competitors or the public. Ultimately, this carries not only risks to intellectual property but also for AI safety. Access to samples of post-training data can be used to create adversarial prompts, which undermine the very purpose of post-training regulation (Kumar, 2024).

An alternative view might be to retain the focus on post-training data, but enforce it through self-regulation, where developers implement checks and balances themselves, without requiring external oversight. This is largely in line with what was proposed in a position paper that advocates that LLM regulation should learn from social media regulation (Appel, 2025), a very different framework than the one applied to mass media content we propose here. This would not only mean that model providers retain control over the data, but might also reduce the risks of external interference in post-training when, for instance, the independence of media regulators is compromised.

5.2. The real issues lie in pre-training

While alignment discussions often focus on post-training, one can argue that the root cause of compliance failures lies in pre-training. The presence of harmful or copyright infringing content, for instance, is often traced to massive pre-training datasets with insufficient content filtering mechanisms (Li et al., 2024). While post-training can be circumvented with adversarial attacks and jail-breaking techniques (Kumar, 2024), pre-training offers the foundational way of preventing problematic content from surfacing in model outputs.

5.3. Personalization matters more than post-training

While many communication theories of mass media focus on one-to-many effects, some, such as the uses and gratifications theory (Katz et al., 1973), argue that effects are more

directly tied with the individual and social uses of media than the overarching content that is conveyed. Following this argument, the personalized interactions that users have with language models condition effects, and therefore elicit regulation to a larger extent than post-training content. This is implied in current regulatory approaches, such as the European Commission's AI act (European Parliament & Council of the European Union, 2024), which forbid specific uses for general purpose AI models, but do not act upon post-training per se.

6. Call to action

Achieving a balanced, fair, and effective regulation of alignment data as media content requires engagement from machine learning, media, and legal scholars.

Until data vaults are widely implemented, machine learning scholars and practitioners should publish editorial statements or guidelines, akin to those published by newsrooms, sharing the values, principles and priorities that guided post-training efforts with a broader audience. This is not done from a technical perspective, but from an editorial content perspective that considers the expected uses and restrictions of the model. Research should be conducted on the technical viability of tiered post-training data sharing and auditability of this data, exploring differential privacy mechanisms, building on tracing solutions such as OlmoTrace (Liu et al., 2025), or improving metadata and tagging of post-training examples to enhance discoverability without requiring inspection of the full post-training dataset. As a first pilot, we suggest that model developers make publicly available a limited and curated dataset of post-training examples that are representative of their main alignment priorities and choices to encourage public discussion. This should not be so extensive as to compromise IP and facilitate jailbreaking of alignment measures, but also not so vague and generic that it limits meaningful discussion.

Media scholars should apply media research methods to publicly available post-training content to inform both regulation and post-training data creation. These can be quantitative methods, such as using computational content analysis to investigate the presence of confirmation bias in preference pairs, or qualitative methods, such as applying critical discourse analysis to investigate how the Global South is given a voice (or not) in SFT examples. Currently, most media and communication research on LLMs has focused largely on model outputs, often limited to single versions of high profile models such as Llama (Rettenberger et al., 2025). If post-training data is to be regulated as media content, an active engagement of scholars working on topics such as representation and media effects is required for an effective collaboration with both media regulators and machine learning practitioners. This ensures that regulation aims at tangible, beneficial outcomes for LLM audiences. While proprietary post-training datasets may be difficult to access, there are many publicly available post-training datasets on repositories such as HuggingFace, ready to be studied with traditional quantitative and qualitative media and communication research methods.

Finally, legal experts and media regulators should publish guidance and examples of post-training data for high profile events such as elections or addressing problematic and nuanced content such as hate speech. Legal experts should explore the transferability of media regulations to post-training data, accounting for added layers of nuance such as data protection, model provider competitiveness, and accountability. Given the current context where the extent to which large language models should be regulated is highly contentious (Judge et al., 2025), serious thought should be given to how these regulations relate to existing regulatory frameworks such as the AI Act and GDPR in the EU, aiming for an outcome that provides certainty and clarity to both model providers and audiences, and does not become an unfair burden to each of the parties. One actionable measure would be to limit the extent to which data labelers and crowdworkers can be legally restricted by Non-Disclosure Agreements from commercial companies, as their perspectives and experiences may be critical for both researchers and regulators investigating pluralism in relation to Large Language Models. This again mimics mass media protections that anonymous sources have in journalism, where their ability to disclose information that is relevant to the public interest without fear of repercussion is a cornerstone of democratic pluralism.

Acknowledgements

Funding: The author gratefully acknowledges the support of the VENI program, which is financed by the Dutch Research Council (NWO) under the grant (VI.Veni.221S.154; "Beyond accuracy: developing a social science approach to assess and improve machine learning data" PI João Gonçalves). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

The author also thanks the reviewers of the Pluralistic Alignment Workshop at ICML 2026 for their thoughtful feedback and suggestions on an earlier version of this manuscript.

Impact Statement

The societal impacts of the proposed position are outlined in the *Societal impacts of post-training regulation* section.

References

Aday, S. The framesetting effects of news: An experimental test of advocacy versus objectivist frames. *Journalism & mass communication quarterly*, 83(4):767–784, 2006.

Agence France-Presse. Google and AI startup to settle lawsuits alleging chatbots led to teen suicide, January 2026. URL https://www.theguardian.com/technology/2026/jan/08/google-character-ai-settlement-teen-suicide. Accessed: 2026-01-26.

Ahmad, A. K., AL-Jalabneh, A. A., Mahmoud, A., and Safori, A. Covid-19 and the resurgence of the hypodermic needle theory applicability in times of crises. In *International Conference on Business and Technology*, pp. 1423–1436. Springer, 2022.

Appel, R. E. Position: Generative AI regulation can learn from social media regulation. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=fRk0nKLKrJ.

Autoriteit Persoonsgegevens. Ran special: Ai chatbots as voting aid. Report ai & algorithms netherlands (ran), Department for the Coordination of Algorithmic Oversight (DCA), The Netherlands, oct 2025. Special Report, October 2025.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., Stosz, C., Sastry, G., Cooper, A. F., Balston, G., Adler, S., et al. Frontier ai auditing: Toward rigorous third-party assessment of safety and security practices at leading ai companies. *arXiv preprint arXiv:2601.11699*, 2026.

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., et al. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655, 2024.

Ceron, T., Nikolaev, D., Stammbach, D., and Nozza, D. What is the political content in llms' pre- and post-training data?, 2025. URL https://arxiv.org/abs/2509.22367.

De Choudhury, M., Pendse, S. R., and Kumar, N. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*, 2023.

Edwards, H. T. and Berman, M. N. Regulating violence on television. *Nw. UL Rev.*, 89:1487, 1994.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). *Official Journal of the European Union*, L(2024/1689): 1–144, July 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj. ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

Eurostat. 32.7% of EU people used generative AI tools in 2025, dec 2025. URL https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251216-3. Eurostat News.

Fan, R.-Z., Wang, Z., and Liu, P. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *arXiv preprint arXiv:2507.16812*, 2025.

Fengler, S., Eberwein, T., Alsius, S., Baisnée, O., Bichler, K., Dobek-Ostrowska, B., Evers, H., Glowacki, M., Groenhart, H., Harro-Loit, H., et al. How effective is media self-regulation? results from a comparative survey of european journalists. *European journal of communication*, 30(3):249–266, 2015.

Gerbner, G. Cultivation analysis: An overview. *Mass communication and society*, 1(3-4):175–194, 1998.

Iyengar, S., Peters, M. D., and Kinder, D. R. Experimental demonstrations of the "not-so-minimal" consequences of television news programs. *American political science review*, 76(4):848–858, 1982.

Judge, B., Nitzberg, M., and Russell, S. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, 44(1):85–97, 2025.

Katz, E., Blumler, J. G., and Gurevitch, M. Uses and gratifications research. *The public opinion quarterly*, 37(4): 509–523, 1973.

Klayman, J. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

Kumar, P. Adversarial attacks and defenses for large language models (llms): methods, frameworks & challenges. *International Journal of Multimedia Information Retrieval*, 13(3):26, 2024.

Lai, H., Liu, X., Gao, J., Cheng, J., Qi, Z., Xu, Y., Yao, S., Zhang, D., Du, J., Hou, Z., et al. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2771–2791, 2025.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Li, H., Deng, G., Liu, Y., Wang, K., Li, Y., Zhang, T., Liu, Y., Xu, G., Xu, G., and Wang, H. Digger: Detecting copyright content mis-usage in large language model training. *arXiv preprint arXiv:2401.00676*, 2024.

Liu, J., Blanton, T., Elazar, Y., Min, S., Chen, Y.-S., Chheda-Kothary, A., Tran, H., Bischoff, B., Marsh, E., Schmitz, M., et al. Olmotrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 178–188, 2025.

McCombs, M. E. and Shaw, D. L. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2): 176–187, 1972.

McLuhan, M., Gordon, W. T., Lamberti, E., and Scheffel-Dunand, D. *The Gutenberg galaxy: The making of typographic man*. University of Toronto press, 2011.

Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., and Fletcher, R. *Digital news report 2025*. Reuters Institute for the study of Journalism, 2025.

Oomen, T., Gonçalves, J., and Mols, A. Rage against the artificial intelligence?: Understanding contextuality of algorithm aversion and appreciation. *International Journal of Communication*, 18:609–633, 2024.

Raeijmaekers, D. and Maeseele, P. Media, pluralism and democracy: what's in a name? *Media, culture & society*, 37(7):1042–1059, 2015.

Rettenberger, L., Reischl, M., and Schutera, M. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17, 2025.

Sarkar, U. E. Evaluating alignment in large language models: a review of methodologies. *AI and Ethics*, pp. 1–8, 2025.

Schrape, J.-F. Social media, mass media and the'public sphere': Differentiation, complementarity and coexistence. 2016.

Shi, Z., Wang, A.-L., Fairchild, V. P., Aronowitz, C. A., Lynch, K. G., Loughead, J., and Langleben, D. D. Addicted to green: priming effect of menthol cigarette packaging on brain response to smoking cues. *Tobacco control*, 32(e1):e45–e52, 2023.

Shoemaker, P. J. and Vos, T. *Gatekeeping theory*. Routledge, 2009.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 46280–46302, 2024.

Tewksbury, D. and Scheufele, D. A. News framing theory and research. In *Media effects*, pp. 51–68. Routledge, 2019.

UNICEF. Ethical reporting guidelines. https://www.unicef.org/media/reporting-guidelines. Accessed: 2026-01-26.

Valkenburg, P. M., Peter, J., and Walther, J. B. Media effects: Theory and research. *Annual review of psychology*, 67 (2016):315–338, 2016.

Van Driel, L. and Dumitrica, D. Selling brands while staying "authentic": The professionalization of instagram influencers. *Convergence*, 27(1):66–84, 2021.

Wall, M. Citizen journalism: A retrospective on what we know, an agenda for what we don't. *Digital journalism*, 3(6):797–813, 2015.

Weeber, F., Ceron, T., and Padó, S. Do political opinions transfer between western languages? an analysis of unaligned and aligned multilingual llms, 2026. URL https://arxiv.org/abs/2508.05553.

Winata, G. I., Zhao, H., Das, A., Tang, W., Yao, D. D., Zhang, S.-X., and Sahu, S. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *Journal of Artificial Intelligence Research*, 82:2595–2661, 2025.

Zeldin-O'Neill, S. It's a crisis, not a change: the six Guardian language changes on climate matters. https://www.theguardian.com/environment/2019/oct/16/guardian-language-changes-climate-environment, oct 2019. Accessed: 2026-01-14.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wang, G., et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.