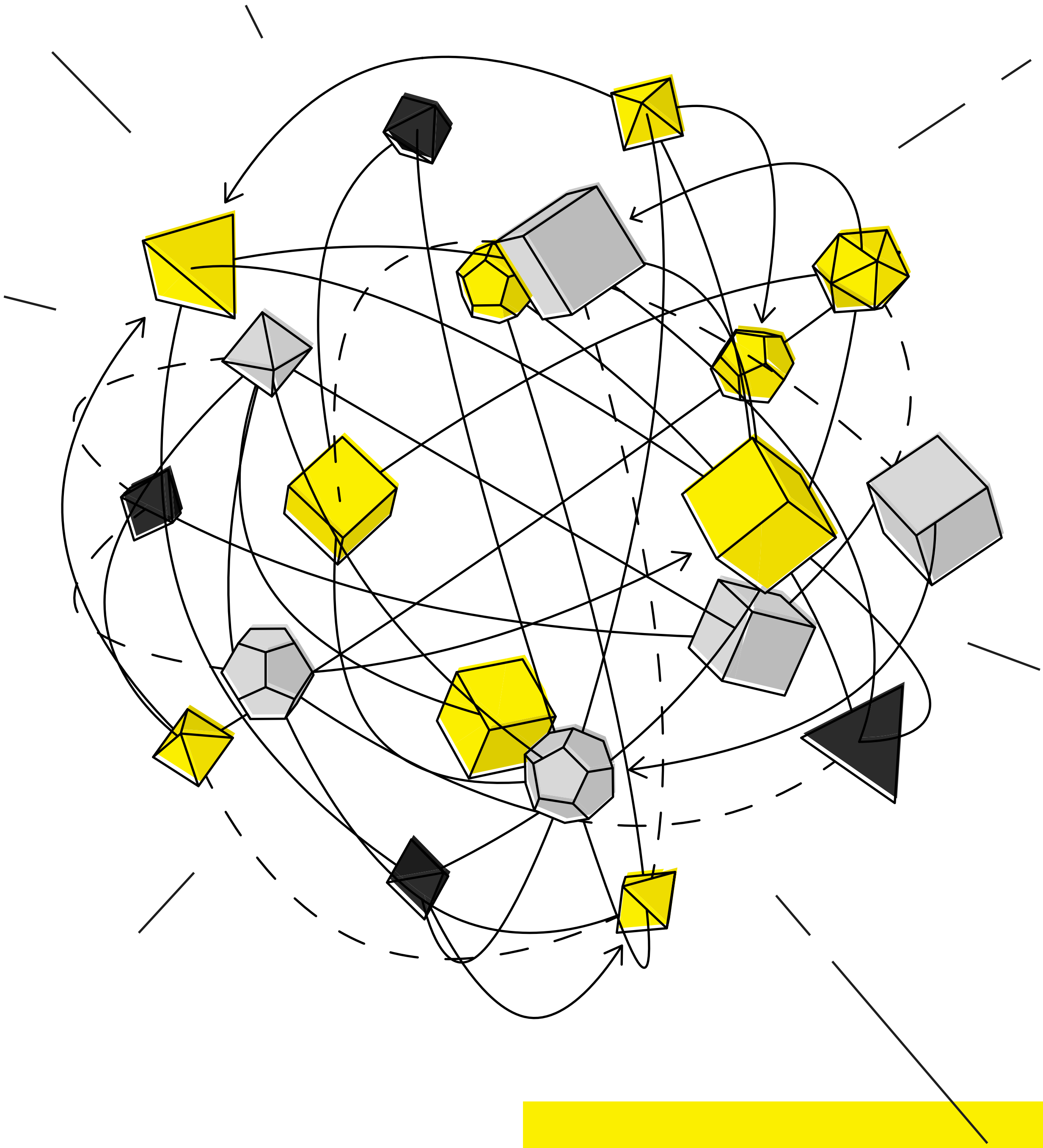
BEYOND AI & COPYRIGHT

funding a sustainable information ecosystem



OPEN
_FUTURE

JUNE 2025

Paul Keller

INTRODUCTION

This white paper explores the impact of the increasing deployment of Artificial Intelligence (AI) in the digital information environment on the sustainability of human information production.

More than two years into the generative AI era, some of the macro-level challenges and opportunities posed by this rapidly evolving technology are beginning to come into clearer focus. In this context, "macro" refers to the structural and societal dimensions of AI's impact—marking a shift away from the individual-level concerns that have dominated much of the public discourse.

While debates have often centered on whether AI will replace creative labour, whether individuals should be allowed to opt out of data use, or whether automation will lead to widespread unemployment, this paper takes a different approach. The author deliberately sets aside speculative concerns, such as existential risk—of which he is highly skeptical—and broader human rights implications. This is not because they are unimportant, but because they fall outside the scope of this analysis.

Instead, this paper focuses on the impact of generative AI—particularly foundation models—on the (digital) information ecosystem and economy. The arguments developed in this paper primarily concern Large Language Models (LLMs) and related approaches that are driving much of the discourse around AI policy and the impact of AI on society.

Specifically, it asks: What does this transformation mean for how we structure access to information? And how does it affect the long-term sustainability of information production in a society that depends on a healthy, diverse, and accessible knowledge commons? As part of this, the paper will develop four key points:

First, the impact of AI models on how humans produce, access, and consume information is still underestimated, especially when it comes to control over information (the information policy dimension) and the commodification of the information commons (the economic sustainability dimension).

Second, there is an urgent and essential need for public AI models and infrastructures—that is, AI models and infrastructures deployed through public funding and governed in the public interest.

Third, in order to understand the impact of AI on sustainable information production, it is necessary to adopt a broad understanding of the information ecosystem—one that includes the wide range of actors contributing to the digital commons.

Fourth, ensuring economic sustainability requires redistributive mechanisms beyond copyright law—specifically, a levy system activated when commercial AI services are deployed, channeling revenue back to all contributors to the information commons.

1. AI IS A NEW SOCIAL & CULTURAL TECHNOLOGY MEDIATING ACCESS TO INFORMATION

In January 2025, Yann Le Cun, Meta's Chief AI scientist, took to LinkedIn to proclaim that :

> *Every institution, library, foundation, cultural group, and government around the world that possesses cultural content should make it available for training free and open AI foundation models. Free and open AI systems will constitute the repository of all human knowledge and culture.*

This observation caused considerable backlash, mainly because this position seems blatantly self-serving for a company that claims that its flagship Llama series of models is open source. It signals an astonishing amount of self-entitlement of one of the most powerful companies in the world. But this paper does not intend to criticize Meta or to adjudicate the complicated question of when AI models should be considered to be open source. Instead, it takes the second part of Le Cun's observation as a starting point: What does it mean if "AI systems will constitute the repository of all human knowledge and culture"? How does this affect the overall information ecosystem, and what can democratic societies that rely on a pluralistic and healthy information ecosystem do to prepare for such a possibility?

If we take this idea seriously—that AI systems could become the repositories of all human knowledge and culture—then it quickly becomes evident that this vision carries profound implications for how we understand AI. We are dealing with a new kind of cultural and social technology that mediates access to information. As observed by Henry Farrell et al in Science,

> *large models should not be viewed primarily as intelligent agents but as a new kind of cultural and social technology, allowing humans to take advantage of information other humans have accumulated. [...] They are analogous to such past technologies as writing, print, markets, bureaucracies, and representative democracies.*

This perspective moves well beyond the familiar narratives of innovation and disruption that focus on the transformation of creative production and individual business models. Instead, it challenges the very foundations of how societies produce, preserve, structure, and access knowledge and information.

AI, in this context, is not simply a tool that enhances the efficiency or reach of existing institutions. Rather, it presents the possibility of displacing them. Foundation models trained on vast troves of publicly available information and other cultural content have the potential to substitute the roles traditionally played by media institutions, libraries, archives, broadcasters, and museums.

These models do not just organize information—they mediate access to knowledge, simulate authority, and generate meaning. Once all publicly available information, including the collections held by these institutions, has been ingested into the training datasets of large-scale

AI systems, there will likely be diminishing incentives—both commercial and user-driven—to access knowledge through the original institutions themselves.

Yet, paradoxically, we may come to rely on these institutions even more—not as primary interfaces to information, but as guardians of provenance, context, and veracity: functions that generative models, by design, are ill-equipped to provide. In this new configuration, media and public knowledge institutions risk being hollowed out: valued for their data, yet bypassed in their role as trusted intermediaries. This calls for a fundamental reconsideration of how we govern the relationship between generative AI and the institutions that underpin the information commons on which these models are built.

Against this background, it is important to acknowledge that the disruption caused by the emergence of AI as a new cultural and social technology mediating access to information cannot —and should not—be taken as a reason to shut it down. As Benjamin Bratton observes, we should

> *[...] presume that the net benefit of widely-available free or inexpensive generative models of the whole of human knowledge including Art as a medium through which new knowledge is made, including Art, is very obviously a net benefit [for society].*

Recognizing that we are dealing with a new social and cultural technology requires us to develop a framework of how they should be deployed and governed in the public interest.

From Public Information to Public AI

The development of generative AI systems is, at its core, an economic enterprise built on the appropriation of public resources. These models are trained on vast amounts of publicly available information—text, images, code, and more—largely drawn from the information commons that is the open web. Or as Robin Sloan puts it in a recent essay, they are only possible because they have been trained in the entirety of the world's publicly available information:

> *The entire internet, or close enough. The whole browsable commons of human writing. From here on out, we'll call it Everything, which is short for Everything That Can Be Accessed Online, Plus Some Extra Pirated Stuff, Probably.*
>
> *In doing so, they hoover up the shared outputs of decades of human expression, repackage this material through statistical abstraction, and reintroduce it into the economy as proprietary services. This cycle—appropriation, transformation, and resale—raises urgent questions about the structure of the information economy and the sustainability of cultural production.*

Seen in this light, all AI is public AI—not in the sense of ownership or governance, but in terms of input value. The utility and power of these models are directly derived from the richness, diversity, and openness of public information and expression. This includes the vast reservoir of online discourse, the digitized holdings of libraries and archives, and the everyday cultural output of millions of people whose contributions make up the digital commons.

This transformation is not, in itself, something we can—or should—seek to prohibit. The public information commons has always been a foundation for creation and innovation. But we must critically examine the implications of this shift for the broader information ecosystem: Who controls access? Who benefits economically? And what happens to the institutions that have historically served as stewards of knowledge?

While this paper focuses primarily on the economic implications of this arrangement, it is important to stress that this transformation also raises serious concerns in the domain of information governance.

Private gatekeepers of public information

It is well documented that the current race for ever more powerful AI models is dominated by a small number of well-resourced commercial actors. These are largely the same entities that already dominate online information processing and digital platform economies, leaving only limited room for new entrants.

When access to knowledge is mediated by AI systems controlled by a handful of commercial actors, new gatekeeping structures emerge. These actors are not neutral custodians of information; their systems reflect commercial incentives, introduce epistemic biases, and may reinforce particular ideological or geopolitical perspectives.

If, as Yann Le Cun suggests, AI models are indeed becoming the new repositories of human knowledge and culture, and are likely to become the primary interface through which societies access information, then it is unacceptable that their development and deployment should be left exclusively to commercial entities. There is a societal need for public alternatives to the dominant commercial AI models.

However, the answer to this dilemma cannot simply be to call for "open source" alternatives—as expressed by Yann Le Cun. While this may sound appealing, it is increasingly clear that openness in AI is a contested and slippery concept. Even truly open-source models face daunting structural limitations. As Widder et al. have argued, the openness of AI models does not extend to the infrastructures required to train and run them at scale. The costs of inference and deployment are significant, and the concentration of computing power remains a major bottleneck. As a result, without meaningful public investment and infrastructure support, even well-intentioned open-source AI projects are unlikely to emerge as viable public-facing alternatives.

Taken together, these factors make it reasonable to assume that, absent any policy interventions, the capacity to provide high-quality AI access at scale—whether for public use, institutional services, or general-purpose interfaces—will remain concentrated in the hands of a small number of powerful actors.

This emerging landscape poses two unacceptable risks:

First, the economic risk of enclosure and commodification. By capturing and repackaging the digital public domain, commercial AI providers effectively privatize the collective cultural wealth of the information commons built up over decades. What was once freely accessible becomes a monetized service, gated behind subscriptions, APIs, or platform dependencies. This process simultaneously undermines the economic foundations of information production itself—when users access content through AI intermediaries rather than original sources, the revenue streams that sustain content creators and institutions are severed. We will return to these economic disruptions in the final two sections of this paper.

Second, the information policy risk of centralized control. When access to knowledge and culture is mediated through systems owned and operated by a handful of firms—concentrated in a few countries—societies lose the ability to shape the conditions under which their members learn, communicate, and understand the world. This is a profound democratic challenge. We will address this in the next section.

2. THE NEED FOR PUBLIC AI INFRASTRUCTURE

To avoid these outcomes, we need to ensure the development and availability of Public AI[1]: data, models, infrastructures, and services that are publicly funded, transparently governed, and designed to serve the public interest. Public AI should not merely replicate the capabilities of commercial models but must embody different values—pluralism, accountability, and inclusivity—and support the long-term sustainability of the information commons on which they depend.

If we recognize that generative AI functions as a new cultural and social technology—one that shapes access to knowledge and culture—then it becomes clear that we need institutional frameworks to govern it in the public interest.

We have done this before. The creation of public broadcasters in the 20th century responded to a similar challenge: ensuring that access to information would not be left solely to commercial actors. Public service media did not merely compete with commercial media—they served a different function: to act in the public interest, uphold journalistic standards, and ensure pluralistic access to information.

In contrast, the internet era saw a critical failure to apply similar thinking. The infrastructure of the internet and the online information space was left almost entirely to commercial actors. Platforms that now function as primary information intermediaries—search engines, social networks, content-sharing platforms—took over in the absence of public alternatives. We are now living with the consequences of this decision: information environments shaped by opaque algorithms, optimized for engagement and advertising revenue, and vulnerable to manipulation and capture.

The rise of AI as a new access layer to knowledge offers a chance not to repeat this mistake. Instead of allowing the next generation of information infrastructure to be fully privatized, we

[1] The term Public AI is used here as defined in the 2025 white paper by Felix Rieker, Alek Tarkowski et al., which presents it as a vision for AI systems that prioritize the public interest. The paper advocates for transparent governance, equitable access, and democratic oversight as alternatives to privately dominated AI development.

must begin building a public AI ecosystem: one that is publicly resourced, transparently governed, and aligned with democratic values. This ecosystem should be modeled, in spirit and structure, on the best traditions of public service media—while also recognizing that we are dealing with a fundamentally new kind of technology.

In this context, Public AI should be understood as a digital public good: a system available for public use, developed in the public interest, and governed in a way that reflects the pluralistic needs of the societies it serves. But while the principle is clear, the institutional form remains open. What kind of governance model is appropriate for a foundational technology like this?

Should we look to digital commons like Wikimedia, which operate across borders with a volunteer-driven ethos? Or should we follow the national public broadcaster model, rooted in public mandates and national frameworks? What we need is likely something in between: a new type of institution—an information intermediary built for the scale of the internet but grounded in the legitimacy, accountability, and public mission of legacy public institutions. In the EU context, this would also mean an institution that is pan-European in scope.

Equally urgent is the question of how such institutions should be financed. While some core capabilities (like public capstone models or open datasets) should be treated as public infrastructure and funded through public investment, the ongoing operation of generative AI systems—particularly inference at scale—requires significant resources.

If we conceive of Public AI as an infrastructural service for public institutions and non-profit entities—embedded in education, media, research, and other public-interest domains—rather than (or in addition to) as models that can be freely hosted and run by anyone, then sustainable revenue models become essential. Providing Public AI as a service entails ongoing operational costs that must be covered.

As we will explore in section four of this paper, these costs could be at least partially offset by revenues collected from entities deploying commercial AI services that have been trained on the publicly available information.

3. THE NEED TO RETHINK THE INFORMATION ECOSYSTEM

Returning to the risk of enclosure of the information commons and the commodification of digital commons by generative AI systems, it is clear that these can only be addressed within a broader economic framework. This is not simply a copyright issue. While disputes over the unauthorized use of copyrighted material have dominated public debate—with rightsholders demanding consent, credit, and compensation—the implications of generative AI for the information ecosystem are far more expansive.

A significant portion of the material that fuels these systems is part of the Public Domain, including the voluntary commons, or is not actively managed: it is out-of-copyright, shared under open licenses, or created without any expectation of economic return. This includes publicly funded content, everyday cultural expression, the knowledge infrastructures maintained by public institutions, and digital commons such as Wikipedia. The current generation of AI models is only possible because of this vast reservoir of publicly accessible information built atop the open web.

The central question, then, is how to ensure that these crucial societal functions—information production, maintenance, and accessibility—remain viable in an environment increasingly shaped by generative AI.

To begin with, this means acknowledging a few core realities: (1) the technology is already widely deployed; (2) models have already been trained on most of the publicly accessible content of the internet; and (3) this new paradigm of accessing information is unlikely to be reversed.

It is important to recognize that this development is not inherently negative. There is nothing wrong, in principle, with accessing publicly available information through AI systems. As Molly White observes:

> *And frankly, if we want to create a world in which every single human being can freely share in the sum of all knowledge, and where education, culture, and science are equitably shared as a means to benefit humanity, we should stop attempting to erect these walls. If a kid learns that carbon dioxide traps heat in Earth's atmosphere or how to calculate compound interest thanks to an editor's work on a Wikipedia article, does it really matter if they learned it via ChatGPT or by asking Siri or from opening a browser and visiting wikipedia.org?*

But for this to be a legitimate and sustainable use of technology, we must ensure that human knowledge production is sustainably maintained as a core pillar of the public information ecosystem. It should be in everyone's interest—including that of AI model developers—that this ecosystem does not collapse.

Ongoing investment in knowledge creation is essential not only to cultural and intellectual life but also to the long-term utility and relevance of AI systems. Without a steady supply of new and diverse information, these systems risk stagnation.

The limits of copyright

What we face, then, is a collective challenge that requires a political solution. At its core lies the unreciprocated appropriation of publicly available information—knowledge and culture harvested by model developers and repackaged into commercial products. This appropriation becomes structurally destructive when AI-generated content floods information channels and AI systems increasingly substitute for direct engagement with original sources, systematically eroding the visibility, traffic, and revenue streams that sustain human information production.

Despite growing public scrutiny and an expanding field of legal challenges, copyright law has proven to be an inadequate safeguard. Across jurisdictions with widely varying copyright regimes, the outcomes are strikingly similar: AI systems trained on much the same public corpus, producing outputs that increasingly undermine the business models of established information creators and intermediaries.

A recent wave of licensing deals between large AI companies and prominent rightholders—such as major news organizations and media companies—may signal a shift, but these remain outliers, and they are unlikely to scale.

More importantly, the logic of licensing begins to falter when applied to the full range of material available in the digital commons. Not everything can—or should—be licensed, and the transaction costs are often prohibitive, especially for smaller or individual content producers. While collective licensing mechanisms might offer a more scalable path forward, they remain underdeveloped, fragmented across jurisdictions, and far from operational at the necessary scale. Moreover, they suffer from systemic shortcomings: the lack of copyright registration and transparency makes it difficult to reliably identify the works represented and link them to their sources—data that would-be licensees, especially AI companies, consider essential.

This presents a fundamental dilemma: AI systems rely on a broad and deep pool of public information to remain useful and relevant, yet the producers and stewards of that information are rarely part of the licensing economy. Moreover, even where copyright exceptions provide legal freedoms for AI training and deployment, these exceptions offer no mechanism to ensure the economic sustainability of information production within those permitted uses. The result is a growing disconnect between those who sustain the public knowledge ecosystem and those who profit from its exploitation.

This disconnect is especially pronounced for a wide range of actors whose contributions are vital to the health of the information ecosystem—but who are often absent from current debates, which tend to be dominated by voices representing segments of the creative industries. These include:

- Cultural heritage institutions that invest in the preservation and accessibility of works often outside commercial circulation.
- Open content projects like Wikimedia, which provide freely licensed information to support public access and reuse.
- Scientific researchers and open access initiatives that contribute peer-reviewed knowledge to the public domain—often advocating for diamond open access models that resist commercial capture and prioritize equitable knowledge sharing.
- Everyday information producers who publish blog posts, videos, datasets, or other content through platforms or independent channels—often without commercial ambition.
- Governments and public agencies that release data and information in the service of transparency and the public good.

What binds these actors together is their position within the information ecosystem. For them, producing and making information publicly accessible constitutes a public good—either as a core part of their mission or as a positive externality of their primary activities. This has always lacked robust economic mechanisms for sustainability, but under the conditions of generative AI, both the costs and the consequences of this gap have become impossible to ignore. The massive, uncoordinated scraping of public web resources for AI training has introduced real and rising operational burdens. Maintaining websites, managing traffic surges, and fending off unwanted bots all require resources—yet these providers receive no compensation, and often not even acknowledgment, for their contributions to the training data that underpin commercial AI systems.

In addition, the rise of an AI-mediated paradigm for accessing information poses another risk to these actors: it threatens to render existing information providers obsolete. As people increasingly turn to generative models to access knowledge and culture, institutions that have historically fulfilled this role—public libraries, archives, open-access publishers, cultural heritage organizations—may see their public relevance, and therefore their funding, erode. When access to information is mediated by AI systems, public awareness of where that information originates is often diminished. This affects press publishers—who worry about the viability of advertising-based business models—as much as it does libraries and museums, which depend on public funding often tied to their visibility and perceived societal impact.

From a societal perspective, this is not an acceptable outcome. Yet much of the current debate remains narrowly focused on securing compensation for rightholders of copyrighted works. If we accept the core premise of this paper—that the development of generative AI has relied on the appropriation of the entire public information commons—then we must adopt a broader perspective on what sustainability means for our information ecosystem. The challenge is not simply to protect the rights of individual creators and their creative industry intermediaries, but to ensure the continued vitality of the public infrastructures that support knowledge production, preservation, and access and the public goods that result from them.

What is needed are new concepts, mechanisms, and institutions that recognize public information as a shared resource—and that enable fair and proportionate value flows from those who extract and commercialize this resource to those who create, produce, curate, and sustain it.

4. BEYOND COPYRIGHT: A FRAMEWORK FOR REDISTRIBUTION

The challenge of ensuring sustainable information production in the age of generative AI cannot be solved within the traditional boundaries of copyright law. What is needed instead is a new redistributive framework—one that recognizes the full scope of the appropriation that has taken place and establishes mechanisms to channel value back to the entire ecosystem of information producers, custodians, and distributors whose work makes AI systems possible.

Expanding the Scope of Recognition

Any effective redistribution system must begin by acknowledging that the challenge extends far beyond the narrow class of professional creators and intermediaries who typically benefit from copyright-based compensation mechanisms. The training of foundation models has drawn upon the entirety of the publicly accessible information ecosystem, encompassing not only copyrighted works but also public domain materials, openly licensed content, government data, and the vast repository of everyday cultural expression that populates the digital commons.

This broader understanding demands a correspondingly expansive approach to redistribution. The focus must move beyond compensating individual rightholders towards sustaining the public goods and infrastructures that make the information ecosystem viable. This means supporting the operational capacity of open access publishers, cultural heritage institutions, Wikipedia, and other commons-based projects —ensuring that these actors can continue their mission of producing and maintaining publicly accessible information. Ultimately, strengthening the public information ecosystem as a whole also benefits individual creators who share their work without expectations of a direct monetary return.

Finding the Right Attachment Point

The implementation of such a system requires identifying a clear attachment point—a specific action or circumstance that triggers the compensation obligation. Here, the limitations of existing copyright frameworks become particularly evident.

Current discussions about the use of copyrighted works by AI model developers[2] typically focus on two points in the AI value chain: (1) reproductions of copyrighted works made during model training, and (2) the use of copyrighted works as inputs for trained models. Both involve reproductions that—depending on the jurisdiction—may or may not require permission from rightholders. While it might theoretically be possible to construct licensing models around either type of reproduction, such approaches face significant practical challenges.

In the case of reproductions made for training, it is often extremely difficult to assess the economic value of these reproductions at the time they occur. Models are typically trained well before deployment, when their commercial potential remains uncertain. Moreover, training occurs once, while the resulting model may be used for extended periods—making it difficult to establish appropriate compensation levels based on training activities alone. Finally, techniques such as distillation enable the training of foundation models on the output of previously trained models, breaking a direct link between the works used to train the original model and the distilled model based on it.

When it comes to the use of copyrighted works as inputs for trained models as part of inference, a different problem emerges: While there seems to be increasing willingness of AI model developers to make licensing deals, such deals are currently focussed on a relatively narrow

[2] Questions about the copyright status of model outputs raise a separate set of legal and policy issues and fall outside the scope of this white-paper.

class of rightholders that can provide access to broad in-demand repertoires such a major news publishers and other entities controlling commercially interesting catalogues. It is unlikely that such licensing arrangements will scale towards smaller rightholders or even cover the full range of contributors to the information ecosystem. This means that relying on inference-time licensing will likely lead to further concentration in the media sector, and thus run counter to policy objectives aimed at supporting a diverse and resilient information ecosystem.

Taken together, these limitations point toward a different approach: attaching the redistribution mechanism to the deployment of trained models in the market, rather than to their training or the use of copyrighted work by trained models (although the latter can exist as a supplementary revenue stream).

A Market-Deployment Levy Framework

Under such a framework, a remuneration obligation would be triggered **when an AI system or other type of service based on models trained on publicly available information is made available on the market**. At this point, the provider of the service would be liable to pay. This could take the form of either a fixed fee based on the number of users or subscribers (a levy) or a percentage of revenue generated by the service (a tax). Each of these approaches has slightly different strengths:

- **A levy** would be more predictable for businesses and easier to calculate. There is also a more direct connection between usage and contribution, and there is ample precedent for such an approach in copyright/media policy. The most obvious disadvantage is that there is a barrier to entry (although this can be mitigated by grace periods).
- **A tax**-based approach has the advantage of being more progressive (larger players pay proportionally more) and that it captures value that is being generated rather than just usage.

Given the clear purpose of the overall intervention—to ensure the sustainability of the information production ecosystem—a levy-based approach seems preferable to a tax-based approach (tax revenue generally flows into the general budget, and is thus harder to earmark and distribute). Such a levy-based approach ("a general payment obligation on providers of generative AI systems") has been previously proposed by Prof Martin Senftleben, although his proposal is primarily focused on the remuneration of individual rightholders[3].

[3] In a similar vein, Valérie Laure Benabou (2023), in From the Three-Step-Test to the Domaine Public Payant, argues that AI outputs—which she considers to be part of the public domain—should be managed under a Domaine Public Payant scheme, functioning as a 'compensatory' mechanism to 'rebalance conditions in the markets concerned.' Distinct from Senftleben or the proposal made here, she suggests that the 'creators' of AI outputs (i.e., AI companies or users of AI systems generating new public domain works) should pay a 'tithe' for creating them—not for using public domain works, but for the act of producing new public domain works that compete with copyright-protected ones.

The proceeds from the levy (or tax) would then need to be distributed among the wide range of information providers, distributors, and custodians whose contributions make AI systems possible. There are four broad classes of potential recipients:

- Creators and rightholders. Distribution to this group could be channeled through existing collective management bodies and based on the size of their existing oeuvre or repertoire.
- Public service media organizations that produce journalistic content and other forms of information. Distribution here should contribute to operating costs, supplementing existing funding arrangements for public service media.
- Open access information providers, digitization projects, cultural heritage institutions, and other entities involved in the preservation and dissemination of knowledge and culture. As with public media, distribution should support their operating costs.
- Public AI models and services. A portion of the levy revenue should be directed toward maintaining these, creating a sustainable funding mechanism for the public AI infrastructure discussed in Section 2 above.

Further modelling and research will be needed to establish both the amount of such a levy[4] and the distribution shares between these three categories of recipients, as well as the allocations within each of the groups. This will include the question of the jurisdictional scope of such a mechanism, which is unlikely to work in individual jurisdictions. This presents a collective action problem, which—given the lack of global coordination mechanisms in this space—ultimately points towards the EU as the most likely candidate for implementation.

Beyond Compensation: Preventing Digital Rent Extraction

This proposed model represents a radical departure from treating AI-related compensation as an issue that can be solved within existing copyright frameworks. While it might still rely on elements of that framework—particularly collective management organizations—for implementation and distribution, its underlying logic is fundamentally different.

The system would operate on the principle that commercial AI services benefit from and should contribute to the commons of publicly available information, regardless of the copyright status of individual components within that commons. Collective management organizations would ensure appropriate distribution of levy proceeds to professional creators and other rightholders, but the system would simultaneously recognize and compensate the broader ecosystem of contributors who fall outside traditional copyright frameworks.

[4] Ideas about imposing levies are not constrained to academic discourse. In a recent interview with Axios, Dario Amodei, the CEO of Anthropic suggested a "token tax" that would kick in every time someone uses a model and the AI company makes money, perhaps 3% of that revenue "goes to the government and is redistributed in some way."

Revenue from the levy should also be used to fund a viable public AI infrastructure that can serve as a check on the power of commercial AI model providers. As argued in Chapter 2 above, this is necessary to prevent a small group of commercial AI model providers from becoming even more powerful information gatekeepers than they already are.

This concern about gatekeeping power becomes even more pressing when we consider the broader structural transformation that AI companies are actively pursuing. The current "race to AI" can be understood not simply as technological competition, but as a contest to become the next landlords of the internet—entities that control access to information and extract rent from that control. A redistributive framework that funds public alternatives is therefore not just about fairness to content creators—it is about preventing the emergence of a new class of information gatekeepers who would have the ability to extract rents from the entirety of human cultural production.

Implications for Licensing and Market Structure

Implementing such a broad redistributive framework brings some challenges to existing conceptions of licensing copyrighted works for AI training. This approach would have several significant implications for how AI development and deployment are structured. Most notably, it would mean that rightholders could not separately license the use of publicly accessible works for AI training purposes. Their ability to issue individual licenses would be limited to non-public or paywalled content—such as proprietary databases, subscription services, or other commercially restricted information.

This limitation is not necessarily restrictive. There remain considerable opportunities to build business models around the production and curation of proprietary data and information specifically for AI applications. As recent developments in specialized data partnerships demonstrate, there is significant commercial potential in creating high-quality, curated datasets for specific use cases.

While this model may be most appropriate for sectors that produce informational content—such as text and images, which are often created for public access and reuse—it may be less appropriate for domains like music[5] and film. These sectors tend to have more complex distribution models, and creators may have stronger claims to retain control over if and how their works are used in AI development.

To account for this, the framework could potentially be combined with opt-out mechanisms similar to those currently in place in the EU, allowing rightholders to withdraw from the

[5] In the music sector, collective management organizations occupy a much more established position than in many other domains. One of the more interesting developments in this context is the intention of the German authors' society GEMA to issue Extended Collective Licenses for "The rights to use musical works (composition and lyrics) for training artificial intelligence systems or models (including generative artificial intelligence), as well as the use of these works or parts thereof or access to them when generating output by the end user."

collective system and maintain individual licensing rights[6]. However, the system would likely be most effective in a scenario where all uses of copyrighted work for training AI models prior to market deployment are broadly permissible, removing the complexity of opt-out mechanisms and creating a clearer, more predictable environment for both AI developers and information producers. Such an approach has recently been proposed by the CREATe Centre in response to the UK government consultation on copyright and AI. By allowing extensive experimentation and innovation during the pre-commercial phase, while establishing clear obligations once systems enter the market, the framework would balance the needs of AI developers with the sustainability requirements of the information ecosystem.

Justifying the Trade-off

This approach represents a significant trade-off: individual control over the use of publicly available content in exchange for broader, more systematic support for the information ecosystem as a whole. The justification for this trade-off lies in the recognition that the current paradigm—where commercial AI systems appropriate the entirety of public information without compensation—is ultimately unsustainable.

By establishing a clear, predictable mechanism for channeling value back to information producers and custodians, this framework would help ensure the continued vitality of the knowledge commons on which AI systems depend. It would also provide a foundation for public AI infrastructure, creating genuine alternatives to purely commercial models.

This may not be a perfect solution, but it is a necessary one. As generative AI continues to reshape how societies access and interact with information, we need mechanisms that recognize both the value of this technology and the importance of maintaining the diverse, sustainable information ecosystem on which this new cultural technology ultimately depends.

5. BUILDING A SUSTAINABLE AI INFORMATION ECOSYSTEM

The analysis presented in this paper points to an urgent necessity to intervene on two interconnected challenges to ensure a sustainable information ecosystem in the age of generative AI.

The first challenge concerns the creation of a public AI ecosystem that enables AI-mediated access to information through systems operating beyond purely commercial logic. This requires substantial investment in institutions and infrastructures capable of providing public AI services at scale, guided by public interest objectives rather than profit maximisation. A critical component of this endeavour will be developing viable economic models that can sustain the development, provision, and operation of such AI systems over the long term.

[6] This is another reason why such an opt-out system would need to be paired with a public registration component to ensure that the distribution of the levy remains equitable and fair.

The second challenge involves establishing a remuneration system that ensures AI models and services trained on publicly available information harvested from the information commons contribute meaningfully to the costs of information creation, curation, and preservation. Such a system must also help finance the public AI information infrastructure described above.

The urgency of implementing this framework cannot be overstated, as the impact of generative AI on various aspects of our information ecosystem is already tangible and set to intensify significantly.

While this paper advocates for an approach that considers the interests of a much broader range of actors than traditional copyright frameworks typically encompass—extending beyond creators and rightholders to include the wider public interest—the upcoming review of the EU copyright framework presents a timely opportunity to fundamentally reconsider how societies can ensure sustainable information ecosystems under the conditions imposed by generative AI.

This paper opened with an observation from Yann LeCun, a senior figure responsible for driving the current race among a small group of entrenched big tech companies to dominate AI technology. It seems fitting to close with another quote from an executive in a similarly influential position. In a recent interview, Kevin Scott, Microsoft's CTO and one of the key enablers of OpenAI, remarked:

> *I don't know whether the deal of the old web is up, but it's kind of time for us to be thinking about some new deal, I think. And I think as we're all collectively thinking about the new, we should do what every good architect does when they're thinking about the new. It's like, what has worked and what hasn't worked for all of the constituents and stakeholders over recent years, and let's try to go make something better that works for everyone.*

This serves as a crucial reminder that we left fundamental elements of the last transformative social and cultural technology—the internet—largely to commercial actors, thereby undermining much of the fabric upon which our democracies are built. We cannot afford to repeat this mistake. This means that we must embed both public interest considerations and fair distribution mechanisms into the very architecture of the next generation of technology poised to reshape the information landscape. The stakes are too high to allow market forces alone to determine the future of how our societies will access, process, and share information.

ABOUT OPEN FUTURE

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture.

Paul Keller is a co-founder and director of policy at Open Future. His work focuses on the intersection of copyright policy and emerging technologies. He works on policies and systems that improve access to knowledge and culture and protect the digital public sphere.

European
Artificial Intelligence
& Society Fund

This publication was made possible with support from the European AI & Society Fund. Open Future is solely responsible for its content, which may not necessarily reflect the views of NEF or the Partner Foundations.