APRIL 2025

JAKOB MÖKANDER
AMANDA BROCK
MICK GRIERSON
KEVIN LUCA
ZANDERMANN
JOSEPH BRADLEY

# Rebooting Copyright: How the UK Can Be a Global Leader in the Arts and AI

TONY BLAIR
INSTITUTE FOR
GLOBAL CHANGE

# Contents

01

# Foreword

The global race for cultural and technological leadership remains wide open. Countries that embrace change and harness the power of artificial intelligence in creative ways will set the technical, aesthetic and regulatory standards for others to follow.

Consider the arts first. Another revolution in media and communication is underway. After the printing press, gramophone and camera, AI is set to disrupt how textual, visual and auditive content is created, distributed and experienced. AI will usher in a new era of interactive and bespoke works, as well as a counter-revolution that celebrates everything that AI can never be. Far from heralding the end of human creativity, AI presents new ways of being original.

The same AI revolution disrupting the creative industries is impacting all areas of society. Scientists use AI to discover in a matter of hours what once took years, health-care providers use it to analyse X-ray images and emergency services use it to find houses damaged by earthquakes. And this is only the beginning. Fuelled by advances in computing power, data, model architectures and access to talent, future AI systems will become increasingly capable.

The government has signalled a commitment to make the United Kingdom a global leader in AI through its AI Opportunities Action Plan, which was announced by Prime Minister Keir Starmer on 13 January 2025. This ambition should be encouraged. If properly designed and deployed, AI can make human lives healthier, safer and more prosperous. Accelerating AI adoption is thus not only a matter of boosting economic growth but also of improving social outcomes.

At the same time, the rapid diffusion of AI across sectors poses urgent policy questions that need to be answered. One of these concerns the data that go into AI training. Today, the application of UK copyright law to the training of AI models remains contested. Unfortunately, the current debate is

too often framed as a zero-sum game, in which AI developers and rights holders are locked in competition for limited resources. This misrepresents the nature of the challenge and the opportunity before us.

The current situation is unsustainable and a lack of clarity harms all stakeholders. This includes creators, who are not properly remunerated for their labour; rights holders, who struggle to exercise control over how their works are used; AI developers, who face hurdles when it comes to training AI models, which limits job-growth potential; and society at large, missing out on benefiting from AI diffusion and adoption. Bold policy solutions are needed to provide all parties with legal clarity and unlock investments that spur innovation, job creation and economic growth.

Critically, the AI revolution presents opportunities as well – not least for creators. From podcasts to filmmaking, AI is already being used to break new ground. The notion that art is threatened by a new tool echoes debates from the past. From the printing press and the Statute of Anne to the internet and music streaming, every technological innovation challenging the status quo has been met with end-of-times claims. But in each case, human ingenuity has prevailed and society has adapted. The progressive solution is not about clinging to copyright laws designed for an earlier era but allowing them to co-evolve with technological change, to remain effective in fulfilling their regulatory and protective aims in the age of AI.

The UK government has proposed a text and data mining exception with the possibility for rights holders to opt-out. This is a good starting point for balancing stakeholder interests, as it would enable AI development in the UK while giving rights holders increased control of how their data are used. But the proposal comes with significant implementation and enforcement challenges with legal, technical and geopolitical dimensions. In this report, the Tony Blair Institute for Global Change (TBI) assesses the merits of the UK government's proposal and outlines a holistic policy framework to make it work in practice.

The UK can be home to both cutting-edge AI development and a flourishing creative sector. Robust opt-out mechanisms and enforcement alongside increased transparency from AI developers and users will help build trust

with creators and rights holders. Additionally, clarifying what qualifies as human creativity and supporting measures to identify AI-generated content will allow society to continue celebrating and rewarding human expression while embracing transformative technological tools.

As with all technological change, there will be disruption. The government should play an active role in helping industries manage abrupt transitions and provide a safety net for individuals. Governments that ignore rights holders' concerns – or fail to provide clarity for AI developers – also face economic, legal and political risks. Yet there are better ways to help creators flourish in the digital age than strict copyright laws. This includes support in honing new skills, accessing compute infrastructure and developing new business models.[1]

Central to TBI's vision is the establishment of a Centre for AI and Creative Industries, a hub that would bring together technologists, artists, policymakers and industry players to forge collaborative solutions and nurture the next generation of creative innovators. This institution would serve both as a symbol and an engine of the UK's commitment to a future where technological progress and creative expression advance in concert.

The stakes could not be higher. The creative industries and the AI sector are central to the UK's prosperity and place in the world, yet their future depends on legislation written before Photoshop was released. Old notions of ownership will not serve rights holders as well as they have for the past 300 years.[2] The government must act now to ensure that the UK remains an attractive location for AI development and investment, while unlocking the vast potential of the country's artistic heritage to shape culture in the 21st century.

**Professor Fernando Garibay, chairman and founder, Garibay Institute**

02

# Executive Summary

The question of how copyright law applies to the training of AI models has rightly attracted much debate.[3] Data are the lifeblood of artificial intelligence, and large language models – such as ChatGPT – have mainly been trained on vast amounts of publicly available data sets containing content scraped from the internet. This has caused friction between rights holders and AI developers. Currently, United Kingdom copyright law provides insufficient clarity for creators, rights holders, developers and consumer groups, impeding innovation while failing to address creator concerns about consent and compensation.

To chart a way forward, the UK government has proposed a text and data mining (TDM) exception with the possibility for creators and rights holders to opt out.[4] Focusing on input data, this would make it legal to train AI models on publicly available data for all purposes, while giving rights holders more control over how they communicate their preferences with respect to AI training. This follows an earlier consultation led by the Intellectual Property Office that explored a similar policy solution but failed to reach a consensus.[5]

In this report, the Tony Blair Institute for Global Change outlines an ambitious programme for cementing the UK's leadership in frontier-AI development and the creative industries. In doing so, we take the government's proposal for a TDM exception with opting out as a starting point. This proposal has come under fire for several reasons. Some question whether the technical tools to enable and enforce the opt-out mechanism are sufficiently robust, while others argue that any commercial TDM exception would cross a red line. While a TDM exception with opt-out will require careful implementation to be effective, we believe it is sound policy for legal, economic and geopolitical reasons.

To begin with, it is important to separate the debates around AI outputs and AI training. AI outputs should not be allowed to reproduce original works without proper licence and remuneration. But prohibiting AI models from training on publicly available data would be misguided and impractical. The

free flow of information has been a key principle of the open web since its inception. To argue that commercial AI models cannot learn from open content on the web would be close to arguing that knowledge workers cannot profit from insights they get when reading the same content.

There are also better ways of supporting the creative industries in the digital age than through restrictive copyright laws for AI-model training. The question is not whether generative AI will transform creative industries (it already is) but how to make this transition equitable and beneficial for all stakeholders. AI is already being integrated into creative workflows, automating routine tasks while enabling new forms of expression. Moreover, the economic impact will vary across sectors and individuals. Rather than fighting to uphold 20th-century regulations, rights holders and policymakers should focus on building a future where creativity is valued and respected alongside AI innovation.

A whole-of government approach is needed to deliver on the AI Opportunities Action Plan (AIOP). The UK government is ambitious about making the UK a global leader in AI, including investment in data, skills, compute and efforts to align government procurement and technology regulation. The Intellectual Property Office's proposal for a TDM exception with opt-out is compatible with the AIOP and would bring UK regulation broadly in line with the EU's. But other jurisdictions, such as Singapore[6] and Japan,[7] have more liberal copyright laws in relation to AI training. A more restrictive policy could potentially push AI developers out of the UK, undermining not only the AIOP but also the government's broader growth agenda.

In truth, geopolitical considerations require urgent and adequate attention. Without similar provisions in the United States, it would be hard for the UK government to enforce strict copyright laws without straining the transatlantic relationship it has so far sought to nurture. The current US administration has indicated that it will not pursue strict AI regulations – and views attempts by other countries to do so as anti-competitive.[8] While there is ongoing litigation in the US around AI training and what constitutes fair use of copyrighted materials, these will be decided on a case-by-case basis.[9] China, the other major AI power, is also pursuing AI development at

breakneck speed.[10] If the UK imposes laws that are too strict, it risks falling behind in the AI-driven economy and weakening its capacity to protect national-security interests.

There is also the question of extraterritoriality. As of today there is a loophole in the international copyright system that allows the copying of training data in a country where this activity is not illegal.[11] Stricter copyright laws in the UK would not prevent publicly available data from being used to train AI models that are then made available in the UK. To close this loophole, the principle of extraterritoriality would have to be applied and copyright rules on models imported from abroad would have to be enforced. But this would significantly complicate AI development and could lead to cutting-edge tools not being accessible in the UK.[12] As a medium-sized nation benefiting from trade, the UK's focus should be on accelerating technology adoption while facilitating the global harmonisation of copyright laws. The TDM exception with opt-out presents a model that can be replicated and consequently enforced.

While the TDM exception is workable, it is only a small part of a bigger story about creativity in the 21st century. Serious support for a transition in the creative industries is essential if the UK is going to continue to nurture its world-leading cultural sector. The path forward lies not in choosing between technology and creativity, but in designing systems where both can flourish in a mutually beneficial ecosystem that respects human expression while embracing AI's transformative potential. This has been done before with digitalisation and online streaming of content – and it can be done again.

Now is the time to act. Based on extensive consultation with rights holders, AI developers, creatives, academic experts and legal scholars, we present four recommendations designed to support both AI development and the creative industries in the UK. The remainder of this report justifies interventions and details their implementation.

# Recommendations

1. **Strengthen AI-preferences standards for rights holders.** The UK government should support internationally harmonised AI-preferences standards developed for an effective opt-out regime. These standards must go beyond the limitations of robots.txt (files used to manage crawler access to a website), offering rights holders greater control over the use of their content while incorporating pragmatic commitments from developers to trace and respect these preferences. Open-source tools can play a crucial role in operationalising these standards, providing a "tools-not-rules" approach that fosters innovation. This was suggested for security management in AI at the recent AI Action Summit, through the launch of ROOST.[13] These tools should also allow be set up to allow rights holders and developers to track content across the web.
2. **Advance a multi-pillar transparency approach.** The UK government should implement policies that include pragmatic disclosures from AI developers, attributional transparency and private regulatory scrutiny. Additionally, work should commence on technological solutions for identifying humanness online, ensuring that rights holders can manage how their content is accessed and paving the way for agentic-AI models that interact responsibly within digital ecosystems.
3. **Establish strong standards for AI and creativity.** To safeguard the creative industries, clear standards must be established regarding creativity and licensing in AI applications. The UK government should introduce a one-off exception allowing major rights holders to license the past 75 years of content for AI training, as recommended in the AIOP. Additionally, standardised contracts between the creative and AI industries should be developed to enable more seamless business partnerships. The UK government must also provide clear guidelines on human creativity, ensuring effective mechanisms exist to distinguish between human and AI-generated works to enable appropriate attribution and remuneration.
4. **Support the transition of the creative sector into the generative-AI era.** The UK government should adopt a proactive approach to supporting the creative sector's transition into the AI era. This can be achieved through targeted funding and the establishment of a new

Centre for AI and Creative Industries (CACI). Given the complexities surrounding technological substitution and the legal implications of the Berne Convention – a treaty setting the framework for copyright law – a dedicated remuneration scheme should be introduced. Strategic investment in AI-ready data for the creative sector will also ensure that the UK remains at the forefront of AI-driven artistic and cultural innovation.

03

# Reframing the AI and Copyright Debate

Generative AI is here to stay. Already, music generator Beatoven.ai has built a fully licensed generative music model,[14] and KL3M has produced a "fairly trained" large language model (LLM) that can perform useful administrative tasks around legal and financial data with only 3.7 billion parameters.[15] As more content falls out of copyright, more will become available to train generative models, even under the most restrictive copyright regimes. Avoiding the introduction of generative AI will delay the deployment of these tools – or, worse, drive consumers to seek the same product overseas.

## Novel Forms of Art Will Emerge From the Generative-AI Revolution

Art has repeatedly adapted to technological pressure and will again in generative-AI era. The diffusion of photography in the 19th century, facilitated by the French government's release of its patent rights in 1839, triggered a profound crisis in modern art yet sparked a series of transformative movements.[16] Impressionists responded by focusing on subjective perception rather than precise representation; by the 1910s, Wassily Kandinsky and Kazimir Malevich had pushed further into abstraction, establishing new philosophical approaches to art. This evolution continued through surrealism and eventually conceptual art, questioning art's materiality.

Similarly, audio-sampling technology in the late 1970s transformed music. Building on the electronic experiments of Karlheinz Stockhausen and Kraftwerk, affordable sampling equipment in the 1980s democratised production. Hip-hop artists turned turntables into instruments; producers created complex soundscapes from multiple samples, leading to the emergence of new genres such as Chicago house and Detroit techno.

Photography and sampling are key examples of technologies that sparked debates about creative ownership but ultimately led to artistic renewal rather than extinction. Artists found new pathways that were complementary or alternative to the capabilities of those technologies.

There are already examples of generative AI and creativity coexisting, with AI being used at the highest levels to help creators. Japan, South Korea, Israel and the US have copyright laws that allow for training and still enjoy thriving creative sectors. *Inside Out 2*, nominated for best animated feature film at the 2025 Oscars, used an AI denoiser from Pixar.[17] *The Brutalist*, another Oscar nominee in 2025, used generative AI for speech editing and visual effects, albeit controversially.[18]

Innovation is inherently unpredictable, making it impossible to predict the future of art. However, policymakers can still create an environment that encourages fair innovation for all stakeholders. The groundbreaking artists of the generative-AI era will likely not be today's celebrated creators, but rather tomorrow's emerging talents who need support from current policymakers to realise their potential. Copyright law alone is insufficient to cultivate a bold artistic future; a more diverse and forward-thinking policy mix is required.

Creators and rights holders are understandably concerned that their livelihoods are at risk in the absence of such a vision. AI models can produce artistic works of a high quality with rapidly decreasing costs in a fraction of the time it would take a human. To add insult to injury, the AI models have been trained on data that includes rights holders' own works, without acknowledgement or any form of remuneration.

## Rights Holders and Developers Disagree on Copyright

Copyright in the UK is a set of economic and moral rights that control how work can be copied and distributed.[19] In an era of printing presses and photocopiers, this was an appropriate tool for rights holders to secure

remuneration for their work and to control where their work was distributed. Yet with the dawn of the internet, old-world copyright is already coming under stress.

Copyright has been used to bring litigation against developers, notably in the New York Times v OpenAI lawsuit in the US, and in the Getty Images v Stability AI case in the UK. The allegations in AI and copyright cases generally split roughly into two parts: first, that the outputs of AI models constitute an illegal copy; second, that using copyrighted works in training data for AI (inputs) is a breach of the image owner's copyright.

There is a consensus that outputs close to an exact copy of the original are a breach of copyright; most chatbots will not print entire chapters of books for exactly this reason. While the degree of similarity that is acceptable is up for debate, this is not a problem unique to AI. Rip-offs are common in film, music and literature, and the legal precedent already exists to help resolve these debates. Copyright is a powerful tool for creators and rights holders in this respect and, as such, we do not focus on the question of outputs in this report.

Using copyrighted material as an AI input is heavily debated. Some countries have adopted a "fair use" policy to permit this, but only if certain conditions are met. For example, in the US, Thomson Reuters recently won a summary judgment against the use of its work in AI training because the end user was a direct competitor. But, in general, copyright controls copying; it does not control other ways in which those engaging with the material might use its intellectual content.

Rights holders argue that when developers use their art to train AI models, the developers should pay for any material from which they make a profit. There are two problems with this argument. First, developers are not set to make long-term profits from publicly trained data. For example, most of the foundational LLMs and image-generation models that people interact with are free at an introductory level and coming under pressure from competitors, as training[20] and inference[21] costs decrease. It is likely that there will soon be little profit to be made from serving the data in LLMs and image-generation models from publicly available data, as the data itself will

be a public good as far as copyright allows. Instead, profit will come from carefully constructed applications and the use of private fine-tuning data, which will be licensed.

In addition, there is no consensus on whether people should pay for everything from which they make profit. For example, many workers in the knowledge economy read the news. They then sell their general knowledge, including what they learned reading the news, as part of their work. They owe no extra money to the newspaper beyond what they might have paid to access it. They might cite the paper, but this is not required by law. Similarly, artists visit galleries, often with no entrance fee, to explore a variety of creative works. Was Tracey Emin expected to reimburse Louise Bourgeois[22] for the transformative experience she had upon encountering her work at the Tate in 1995?[23] The relationship between originality and imitation has always been ambivalent, from classical art to the present day.[24]

Of course, it is reasonable to consider how copyright should work differently for AI training, especially concerning the arts. But this kind of nuance is very different to the claim that all AI training is inherently stealing, and that copyright should be extended to cover the spread of all human ideas through AI. Few would say that any reproduction is stealing and, even within the strict confines of copyright law, courts distinguish between generic ideas and original expressions.

It is hotly debated whether model weights should be thought of as a copy of the training data. Given that individual items in the training data set can, in some cases, be recovered from the final model, it appears that some machine-learning architectures compress their training data into the model weights in some way.[25] This highlights the problem of applying the traditional use of the word “copy” to computers that, by design, make copies of information billions of times a second. To the extent that there are copies of protected works in model weights, the significance of these copies can be mitigated by proper guardrails at the output stage that prevent users from receiving verbatim copies, which would clearly be a copyright infringement.[26]

# Beyond Creators Versus AI: The Wider Implications of Copyright Policy

This debate is often framed as a zero-sum game between artists and AI developers. Considerable media attention has been paid to leading figures in the arts[27] and their interaction with AI. However, the holding of rights extends well beyond the creative industries; copyright has played a huge role in scientific information, for example. The UK government spends tens of billions of pounds per year[28] on research and, like creative outputs, scientific papers are subject to intellectual-property (IP) protection. This creates a perverse system whereby a government pays three times for cutting-edge research: once for the original study, once for the publisher to host the paper and once for the right to use that paper in AI-training data sets. This means that the taxpayer also loses from the current copyright policy.

As the Centre for Regulation of the Creative Economy points out,[29] this was a significant motivation for the government-commissioned Hargreaves review of IP and growth in 2011.[30] But the non-commercial text and data mining (TDM) exception is too narrow to be useful to the NHS and many other bodies which work with commercial suppliers. The inefficiencies of this system and the need for greater use of open-access publishing have been flagged in a report by TBI: *A New National Purpose: Accelerating UK Science in the Age of AI*. Science and innovation will increasingly depend on the ability of computers to search the internet for information.

The debate also extends well beyond generative AI and artists, in that AI includes recommendation algorithms, classifiers, computer vision, network analysis and much more. The law in the UK affects all AI training, and to this extent regulating inputs is a crude approach compared to regulating outputs. This is part of the US fair-use policy, where users of protected material must demonstrate that they do not interfere with the potential market for the work.

Furthermore, rights holders and developers do have mutual interests. Taking the worst-case scenario to its logical conclusion, a poorly executed TDM exception could result in a collapse not just of the creative industry but of AI models as well, as developers struggle to find new content.[31] If models are to adapt to new trends, they will need new data. If there are no creators left due to the economic impact of generative AI, developers will lose out too. By disrespecting creators' preferences, developers could remove a crucial pillar of their own success.

Time is not on anyone's side. Large, effective foundation models already exist and are publicly accessible. They will continue to grow in capability and will be used by an increasing number of people. They will be developed around the world, in jurisdictions with very relaxed copyright laws, and used as tools to the extent that they will inevitably make some jobs redundant. At the same time, countries with restrictive laws prevent developers from starting up where they want – such as in the UK, with its enormous advantages in skills and data sources – and are pushing companies to move to less restrictive countries. The longer governments take to tackle the issue of AI and copyright, the more they will inhibit innovation and entrench large AI developers in the global competition for AI leadership.

Rights holders' and creators' attitudes towards generative AI are not uniform. Independent and/or local journalists may care more about accessing useful AI tools to quickly produce blog posts, podcasts and social-media content; bigger-name, more established writers will likely want to control the diffusion of their columns. An up-and-coming fashion designer could be excited to see that their new style has been picked up on social media and incorporated into generative-AI services (in the hope that demand for the real thing is likely to follow), but an established fashion house might worry that AI could make it easier to produce knock-offs of their designs. When policymakers find themselves asked to protect rights holders, they must ask *which* rights holders, or risk entrenching existing hierarchies between creators and rights holders.

An important point is that creators and rights holders are not always the same thing. While some creators of art retain all the copyright inherent in their output, many creators produce work for companies that own or

administer some of their copyright. For the purposes of this report, we will use these terms interchangeably, except where it is essential to distinguish between them.

Furthermore, this debate connects two very different communities, one of which focuses on human expression (creators), and another that focuses on data (developers). Many in the artistic community strongly object to having their work referred to as data; developers point out that once content is online, it does exist as data – and that the machine-learning process can only see art as data. In this report we will use both "content" and "data" where appropriate.

04

# Understanding the Regulatory Landscape in the UK

There are two copyright exceptions for AI in the UK: the temporary copy exception from section 28A of the Copyright, Designs and Patents Act and the non-commercial TDM exception introduced in 2014. The former allows for temporary copies of "no independent economic significance" when accessing works online. This is designed to facilitate data transfer on the web. If this exception didn't exist, making a temporary copy of a news article to read could be considered a breach of copyright. The fact that the ideas in the copies may ultimately be used to make a profit does not make the copies independently economically significant; many more steps are involved between the copy being made and the knowledge being used. This suggests that section 28A could be used to defend AI training on public data. However, the non-commercial TDM exception, which has no opt-out option, allows AI developers to train models on copyrighted data if there is no profit made. That the Intellectual Property Office felt this was needed suggests that the temporary copy exception does not provide any protection after all. This uncertainty, as well as ambiguity around the definition of "non-commercial", was noted in the Patrick Vallance report on pro-innovation regulation for digital technologies.[32] The effect of this is that little AI training takes place in the UK.

## The UK TDM Proposal Faces Serious Hurdles

In December 2024 the Intellectual Property Office announced that it was considering extending the non-commercial TDM exception to a commercial TDM exception with an opt-out for lawfully accessed works. This is already EU policy, and it has support from the prime minister in the AI Opportunities Action Plan (AIOP). Those rights holders who do not want to be involved in AI training would be able to opt out by stating on their web domains that they do not grant permission for their content to be used in AI training. Again, this would be similar to the EU AI Act, but difficult to implement.[33]

The proposal also faces geopolitical hurdles. While the TDM exception with an opt-out proposal would align closely with the EU – in the short term at least – it would lack some of the flexibilities found in the US (at the AI Action Summit, the US administration made it clear that it will not implement restrictive rules), Singapore, Japan and Israel.[34] As such there is disagreement not only between rights holders and developers, but also between foreign-policy and defence experts who are concerned about how the UK positions itself at a time when it is pivoting towards "hard" definitions of AI safety,[35] as well as the use of AI in defence (see *Reimagining Defence and Security: New Capabilities for New Challenges*).[36] While the creative sector is essential to a flourishing society, this debate must be seen in the context of AI's emerging role in strategic competition on the international stage.[37]

A TDM exception on its own is not enough to build a strong AI ecosystem in the UK. While many developers say that UK copyright is *one* reason for not training in the UK, there are many other factors, including a lack of compute, high energy costs and a lack of talent. A proper AI strategy needs to address all these factors, as laid out in the AIOP and TBI report *State of Compute Access 2024: How to Navigate the New Power Paradox*.

Furthermore, some AI development does take place in the UK, but it often happens on remote servers in distant countries. The benefit of a TDM exception for AI business in the UK is that it would reduce the cost and dependency of needing to train abroad and provide legal certainty for those wanting to build on foundation models from overseas.

This is part of the issue in the Getty Images v Stability AI case, and resolving it in favour of innovation will genuinely support UK developers in bringing AI products to market. The case also highlights the importance of balancing extraterritoriality concerns: the EU TDM exception requires that products sold in the single market obey single-market rules, regardless of where the training happened. This amounts to an import ban on most frontier-AI models. This would be incredibly difficult to enforce given that cutting-edge models can now be run by any consumer with a laptop,[38] and it would hold the rest of the UK economy back technologically while preserving a 20th-century creative economy in the UK. If UK and EU rights holders want to

ensure that competing products (not adhering to TDM exception rules) are not introduced to their markets, there are far more flexible tools (such as fair-dealing provisions) for tackling this than a blanket import ban.

## Governing an Opt-Out Policy Is Challenging

The global nature of the internet demands a unified, worldwide solution. Content hosted by rights holders, regardless of jurisdiction, is inherently accessible for AI training. This fundamental characteristic of the open internet calls for a pragmatic approach. While international collaboration is crucial, it is equally essential for the UK government to maintain independent policy tools and a clear vision for its creative and technological industries.

Any solution must be adaptable across industries and businesses of all sizes. A model that benefits large rights holders in the publishing industry may be entirely unsuitable for smaller scientific organisations, such as those dedicated to pioneering medical breakthroughs or developing solutions to combat climate change. The current UK copyright policy does little to protect rights holders from generative AI (beyond protection from exact copies) while holding back the UK tech sector and favouring AI incumbents. The Intellectual Property Office TDM proposal is challenging but workable if all stakeholders engage with the latest technology and standards.

05

# Strengthen AI-Preferences Standards for Rights Holders

***Recommendation 1:*** *The UK government should support internationally harmonised AI-preferences standards developed for an effective opt-out regime. These standards must go beyond the limitations of robots.txt (see Spotlight, below), offering rights holders greater control over the use of their content while incorporating pragmatic commitments from developers to trace and respect these preferences. Open-source tools can play a crucial role in operationalising these standards – a tools-not-rules approach fosters innovation. This was suggested in the context of security management in AI at the recent AI Action Summit through the launch of ROOST.*[39] *These tools should also be set up to allow rights holders and developers to track content across the web.*

## Implementation Problems with Opt-Outs

The fundamental problem with opt-outs is that the internet is structured around unique resource locators (URLs), but the works protected by copyright are not.[40] For example, a recording of a song encompasses songwriting, performance, recording rights and more. These exist in any copy of the recording. The URL of the copy tells you where the recording is, but not who owns the various rights inherent in that work, nor which of those rights they might have chosen to exempt from TDM.

Furthermore, while opting out will prevent rights holders' work from being used in the future, their work may have been used in some older trained models. Fortunately, old models are rarely used by consumers: for example, the GPT-3.5 model that first powered ChatGPT is unavailable only a couple of years later. Subsequent scrapes for new models will have been respecting the updated robots.txt.

SPOTLIGHT

## How does robots.txt work?

The term “robots.txt” refers to a type of webpage attached to a particular website. It supplies bots (robots) with machine-readable rules about how to behave on that website. Different websites have different rules – and TBI is no exception.

FIGURE 1

## The robots.txt file on the TBI website

```
user-agent: *
allow: /
Disallow: /*?*

sitemap: https://institute.global/sitemap.xml
```

It allows any robot (the first line identifies any bot that finds the webpage) to read any page (the second line).[41]

Ed Newton-Rex [42] is a British composer and former vice-president of product at Stability AI. He takes a different approach.

FIGURE 2

# The beginning of Ed Newton-Rex's robots.txt

```
# Squarespace Robots Txt

User-agent: GPTBot
User-agent: ChatGPT-User
User-agent: CCBot
User-agent: anthropic-ai
User-agent: Google-Extended
User-agent: FacebookBot
User-agent: Claude-Web
User-agent: cohere-ai
User-agent: PerplexityBot
User-agent: Applebot-Extended
Disallow: /
```

This targets several bots and prevents them from accessing any page on the site. The standard that defines how these rules should be written and read comes from the Internet Engineering Task Force (IETF) and is called RFC 9309.[43] Whenever a bot arrives at a website, developers are supposed to programme it to check the robots.txt, interpreting it as per RFC 9309.

Webmasters can try to block bots in general but cannot specify for what purposes bots can access the site. This worked well in an era when most bots were either academic – tracing out the structure of the web – or search engines, for which websites are now generally happy to be indexed. With scraping for AI and retrieval-augmented generation (RAG) taking up an increasing amount of bots' time, it has now become an issue.[44]

Furthermore, large-scale scrapers were once limited to a few major players such as Google and Common Crawl. However, with the rise of AI, the landscape has become far more competitive thanks to numerous entrants.

In the US alone, companies such as OpenAI, Meta, Google, Apple and Anthropic are vying for dominance in the frontier-model space. As of early 2025, new contenders from outside the US, such as DeepSeek and Le Chat, have also emerged. Keeping track of the growing number of bots and their evolving purposes has become increasingly challenging.

In addition, the distinction between crawling and scraping means that there will be work listed in crawling databases without an opt-out, but which is opted out by the time it is scraped for training, potentially months later. Crawling is the act of mapping the internet – working out which sites exist and roughly what kind of content they serve. LAION is a crawl – its data sets are essentially lists of URLs.[45] Scraping is the act of taking those URLs and downloading the content they address to use for AI training. The delay between these two actions presents a problem and needs to be addressed.

Using opt-outs is not simple: robots.txt is a machine-readable file that is neither accessible nor understandable to many. Furthermore, while it is easy to stop all robots in one go,[46] rights holders typically only want to block robots searching for training data or retrieval-augmented generation (RAG) content. This results in having to try and cherry-pick which bots to block, including search-indexing bots that are essential for monitoring and navigating the web,[47] and sometimes multiple bots per company for different products. This is time-consuming and error prone.

Opt-out regimes will be trusted only if developers respect them – and that will require transparency regarding training inputs. Transparency is also relevant beyond the enforcement of copyright law: for example, in understanding how a model interacts with protected characteristics or how it communicates confidence in its outputs.[48]

It is impossible to guarantee the integrity of the copyright of any work on the internet. Once material is on the open internet, it is extremely difficult to prevent it from being copied for any purpose, let alone the specific purpose of AI training. While the internet is incredibly useful for rights holders, no individual (or corporation) has a right to the perfect internet for their economic interests, nor the right to legislate it into being. The extent to which people share content online may change, much as it did when people took greater control of their privacy on social-media sites.[49] This will also mean rights holders will be more careful about what they choose to license if their customers are not going to protect their content downstream. This emphasises the level of mutual interest in the copyright problem, in that no one wants to see web content disappear from the web.

## Making Opt-Outs Useful and Accessible

Accessing the robots.txt file on a website is often difficult. For example, wordpress.org plugins that manage users' robots.txt for them exist,[50] but these processes could be much simpler. Meanwhile, social-media sites often offer no control at all. Large tech companies have a role to play here in educating rights holders on how robots.txt works and ensuring that there are simple, one-click solutions wherever possible for updating robots.txt for non-technical users.

Part of the problem is that the existing standard for robots.txt is unsuited for the age of AI. The Robots Exclusion Protocol (REP) was written by the Internet Engineering Task Force (IETF) and allows users to opt out of either crawling for all bots or one bot at a time.[51] Work on a new AI-preferences protocol is underway and is likely to include ways to customise the kinds of AI use for which material can be applied.[52] This protocol should allow webmasters to filter robots by purpose rather than just by name. This

standard should be developed with and then respected by industry, especially the largest tech companies, and governments around the world. RFC 9309 and its successor are not the only options: the World Wide Web Consortium, which is the main international-standards organisation for the internet, is working on its own protocol, TDMRep.[53] However, robots.txt is already widely used and is likely to remain in use until a solid replacement is suggested.

Domain-level opt-outs will not be granular enough; content-level opt-outs would be needed to correctly communicate the preferences of the relevant rights holders in, for example, a news article. However, this requirement does not exclude, nor reduce the need for, domain-level opt-outs. They can also state that on a certain domain, opt-out checks must be conducted at the content level. For websites that honour content-level technologies such as C2PA[54] or ISCC code lookups,[55] webmasters can specify which form of communication they are using and scrapers can act accordingly.

For the UK's TDM exception to be useful, how it interacts with the existing non-commercial exception should be taken into consideration. First, the new exception should extend to databases as well. This is included in the EU's TDM exception and is an important omission from the non-commercial TDM exception that already exists. Second, if the existing exception is retained, it should be adapted to allow for more scientific research than it currently does. The NHS, for example, struggles to take advantage of the exception because its work crosses many trusts and often interacts with the private sector. Rules specific to health care or biotechnology could be useful here, in the same way that there are specific rules for the education sector.

## Tackling the Diffusion Problem

The second component of a successful opt-out regime is a rule that those using the TDM exception are obliged to check that training material has not been opted out elsewhere by the rights holder; simply following a link to a domain that does not opt out in its robots.txt is not enough. If someone's

work has been copied to a site without an appropriate opt-out, the developer should check that the work located by the link had not been opted out elsewhere.

The tools to do this kind of tracing already exist. AI company Spawning has developed a Do Not Train registry that allows artists to tag work around the internet as copies of their original.[56] Developers can then use "data-diligence" software from Spawning to check whether URLs have been opted out.[57] The check is straightforward – one line of code once the package has been installed and imported – and the service is currently free. Given that the technology (a database and reverse search) is well understood, similar tools are likely to appear. The tools do not need to check every possible copy – they only need to identify the original rights holder and their preferences. Spawning's tools also allow developers to respect other opt-out methods, such as HTML tags[58] and the C2PA standard where those metadata have not been stripped.

Governing these tools will be a challenge. A federation of registries could be created to share data, so that content from different jurisdictions is covered with any interface. Alternatively, one organisation might emerge as the standard, but developers and rights holders will need to agree on where it is based and who runs it. Bodies such as the ROOST initiative will continue to have an important role and need to be supported by tech companies and governments.

Rights holders are also at risk of being overwhelmed by too many standards,[59] so the UK government should coordinate the multilateral adoption of just one standard. Internationally harmonised standards will be much more powerful than a collection, and only with international consensus will countries be able to establish norms around AI training that work for everyone.

There will also need to be safeguards, so that only the true rights holder should be allowed to opt a URL in or out. To some extent this problem has been solved – see ContentID on YouTube – but it is important to highlight

the challenges ahead. Tools would benefit from standards that the government can provide that strike a balance between rigour and competitiveness.

To address the timing problem (work being listed in crawling databases without an opt-out, but which is opted out by the time it is scraped for training, potentially months later), opt-outs should be checked every time data are used in training. Some developers do maintain huge libraries of data – basically copies of the entire internet – to avoid downloading data again. Such a library would amount to an illegal copy in the true sense. The "time-of-training" commitment also guarantees that content used in old training data isn't included in new models if its opt-out status has changed. Developers can maintain copies of content for which they have an explicit licence, to avoid the costs of repeatedly copying.

Checking the opt-out status of a URL at training time will change the crawling ecosystem as well. Large crawlers such as Common Crawl (CC)[60] will be less useful if everyone has opted out because they are worried about copyright abuses. Instead, if rights holders are confident that their rights will be checked at the time of training, they can still allow CC to crawl, trusting that scrapers will skip their content. This highlights the importance of transparency and respect for opt-outs.

Maintaining these registries could require an unreasonable amount of work from rights holders. Fortunately, this problem is mostly solved by existing copyright tools. Google's ContentID system for YouTube automatically checks videos for copyright violations.[61] Copyscape allows users to trace their content as it diffuses throughout the internet and can then be used to add content to websites such as Do Not Train.[62]

Soon it will be technically simple to build AI agents that can track your portfolio, maintain registries and initiate robot-to-robot interactions with other websites, asking them to remove your content.[63] How these tools interact will emerge as demand for them grows, but they are technically feasible. The existence of these agents will also simplify content attribution

for AI companies, enabling them to effectively track online content origins. This eliminates plausible deniability for developers who claim ignorance about opted-out work appearing in their systems.

Respecting and enforcing an opt-out rule is more straightforward than is commonly realised; new standards and technologies are sufficient to maintain a mutually beneficial opt-out regime. But even a well-functioning opt-out regime would not solve all challenges surfaced in the current debate around AI and copyright. For example, effective enforcement also requires improved transparency from both AI developers and users online. In addition, the opt-out regime does not deal with the fact that generative AI is going to seriously change the way that artists work and interact with their audiences. The following chapters address these challenges.

06

# Advance a Multi-Pillar Transparency Approach

***Recommendation 2:*** *A comprehensive multi-pillar approach to AI transparency is essential. The government should implement policies that include pragmatic disclosures from AI developers, attributional transparency and private regulatory scrutiny. Additionally, work should commence on technological solutions for identifying humanness online, ensuring rights holders can manage how their content is accessed and paving the way for agentic-AI models that interact responsibly within digital ecosystems.*

The ultimate means of upholding respect for copyright is the threat of discovery. In the AI-scraping context, this requires a level of transparency from developers. Much as food producers are required to disclose the ingredients of what they produce, it is reasonable to expect AI developers to disclose something about the ingredients of their models. The goals of transparency for AI are also wider than defending copyright, such as concerns about the robustness[64] and ethical provenance[65] of the data.

Transparency is more important an idea than is realised for reaching an agreement between rights holders and developers. If developers deploy strong transparency solutions, more rights holders are likely to leave their content open and developing AI becomes easier. Developers lobbying for looser transparency requirements need to be aware of this trade-off.

There are three styles of transparency that can support copyright. The first is passive transparency, whereby developers simply list what they use. To what level of detail should these lists be made? OpenAI's paper on GPT-3 lists which common databases it uses and leaves it up to users to review the details.[66] The Open Data Institute has called for more such transparency[67] as it can be effective for research projects and proofs of concept. The draft EU AI Code of Practice is forward-looking in this way, stressing the methods of data acquisition and processes used to verify the data.[68]

More challenging are frontier models that have proprietary crawlers and sophisticated data pipelines. At the other extreme to the GPT-3 approach, developers could be forced to publish a database of every URL they have accessed, with dates.[69] However, hosting a database of tens of billions of URLs could be expensive for small AI developers and push the market even further into concentration. While developers may scrape a large amount of data, much of this will not be ingested into the training set or may only be used for testing.[70] This information is vital to AI engineering and is an area of huge competition between AI developers.[71] Furthermore, URL-level transparency would not necessarily make it easy for rights holders to police their work. They would still have to interrogate the databases of tens of thousands of different models, at least monthly, for possible updates.

Programmatic scraping checks of opt-out preferences would be better. Developers write code for the scraping process that checks whether a given URL can be scraped by parsing its robots.txt or similar communications. Google has already released a version of the code it uses.[72] Developers can deploy this logic in a small program on their website so that concerned parties can check at a URL level whether their work could have been scraped by the developer. This allows rights holders to be confident that their opt-outs are being respected, without infringing on the competitive secrets of developers or having to host massive databases of URLs for tens of thousands of developers. It also allows developers to demonstrate what content has not been scraped after any copies are deleted.

The second kind of transparency is attribution transparency, using tools such as Uhmbrella and ProRata. There is a lot of research on extracting training data from deep-learning systems, including text[73] and image[74] models, so it has always been reasonable to expect these attribution tools to work. They can be used to probe models for content that has been opted out; if a model shows a statistically significant number of hits for such content, rights holders can launch a legal case. Alternatively, governments and collective management organisations (CMOs) can investigate how the work of their stakeholders has been used in generative models, consensually or otherwise. These can then inform decisions about remuneration, as discussed later in this report.

The final kind of transparency is regulatory transparency, whereby regulators investigate a developer's code and databases to verify that it is taking the necessary steps to respect opt-outs. This could happen behind closed doors, and developers would be compelled to present the actual code used and ensure that it is well explained, with the steps taken to comply with the opt-out policy highlighted. This method allows rights holders to have confidence that their choices are being respected without having to worry about trade secrets. This option may not suit governments that do not have the capacity to comb through hundreds of thousands of lines of code, or that are happy with a combination of passive transparency and an attribution program. It will also require serious international collaboration, especially in an era of intense technological competition.

The ideal policy response would be a mix of these three pillars, and that mix would have to reflect transparency concerns beyond copyright (including privacy). But there are workable transparency options that can hold developers to account in their use of content and prevent a serious disincentive for abuse, without generating thousands of mostly duplicated URL databases.

## AI Summaries Present Problems About Identity

Google[75] has a unique role in the web ecosystem as the provider of Google Search, which accounts for nearly 90 per cent of search-engine market share.[76] One consequence of this is that being crawled by Google to appear in Google Search results is a necessity for a business that wants to be successful on the web.

Webmasters can opt out of the Google bots that crawl for AI-only applications, such as the Google Gemini chatbot.[77] However, it is not possible to opt out of the crawler that powers Google Search and its AI summary features. News-media outlets are rightly frustrated that summaries of their recent articles appear at the top of many search results, depriving them of the clicks from which they generate revenue. This process, often a type of RAG, is useful for preventing LLMs from hallucinating, but to news

media constitutes an unfair violation of copyright. This is already having an impact on companies that depend on visits to their sites for advertising revenue and new subscribers.[78]

Google is making progress on this, signing deals with copyright boards in Europe[79] (and, more recently, with news organisations such as Associated Press) to license their work.[80] It is important to note that these deals are being made when Google could choose to train on publicly available news content in less restrictive regimes. The legal certainty, technological cooperation and public trust built by these licences is clearly valuable to technology companies. Furthermore, if news-media outlets disappear there will be little quality news for people to search for.

An agreement from Google to separate its search-index bots from its pre-training and RAG bots would be massive for restoring rights-holder trust in tech companies. While it is likely that improvements to AI-preferences standards and agreements between Google and major news providers will continue, getting such an agreement from Google would be one of the most significant non-policy interventions that the government could make. It is essential that developers and news-media groups avoid the toxic collapse that would result from a closed internet.

The debate around IETF standards and the "Google problem" highlights the need for robust ways to prove the intent of internet users. It is widely known that developers employ anonymous third-party scrapers that bypass blocks on IP to pass on scraped data to AI models and search engines.[81] One way to verify intent is by proving personhood, because a person in practice cannot be scraping content for AI training.[82] By "proving personhood" we simply mean some technology to prove to the web owner (for example, the BBC) that you are a *person*. This does not mean proving your identity, which would be a serious privacy concession, but that you are a person and will not therefore be gathering training data.

A white paper with authors from institutions including OpenAI, Microsoft, the University of Oxford and the Massachusetts Institute of Technology has laid out steps for governments to take to move closer to proofs of personhood,

albeit from a privacy perspective.[83] Cloudflare, a major content delivery network, has emphasised the need to develop standards for bot transparency, which proofs of personhood could support.[84]

The World Foundation[85] has already developed hardware and software proposals for proofs of personhood. The hardware for biometric recognition in personal devices (via irises, faces or fingerprints) has been around for some time, even if the protocols are still missing. A future where news websites request that clients verify their personhood (or verify that the client is a device used by a person) is technically feasible and would not materially impact users' privacy.

The need for proof of personhood will become more acute as agentic AI becomes widespread.[86] This creates new challenges and opportunities for legislators. Beyond just respecting opt-outs, legislation may need to be put in place that requires those deploying bots to respect personhood requirements. These could be enforced by biometric methods such as WorldID, dramatically reducing the scope for bot abuse.

The law should also expand to allow contracts to be "signed" between robots. Terms of service on websites that prohibit the use of data for AI training will generally not be visible to crawlers or scrapers and may not be enforceable.[87] Law on smart contracts is already happening[88] and will need to continue to ensure that the behaviour of bots is covered.

This should not be taken as being in favour of opt-outs in natural language. While it is possible to interpret opt-outs in natural language, this system would be computationally expensive and inexact. Norms around respecting copyright cannot exist on such shaky foundations, especially when a proper machine-readable standard has already been demonstrated with RFC 9309.

## Defensive Tools Are a Partial Solution

The discussion so far has focused on standards that well-meaning actors should be expected to follow. However, the internet is full of bad or misinformed actors, against whom robust tools are needed.

There are no defensive tools that are completely effective against copyright attacks when AI preferences as expressed in robots.txt are ignored. This relates to our earlier point that protections for rights holders will come in layered systems, much like an airport attempts to provide safety at multiple points without guaranteeing security at any one point. The tools that do exist will improve as understanding of scraping behaviour, generative AI and the needs of the rights-holder community continues to grow.

Network defences such as Cloudflare's Bot Management system[89] and Spawning's Kudurru[90] protect websites by identifying scrapers by their browsing behaviour, and either block or frustrate the scrapers by returning the wrong content. While the intelligent use of networks of bots can circumvent these defences,[91] a well-aligned actor who has simply misprogrammed their scraper will be detected and stopped.

Attribution tools will allow outsiders to detect training data in the output of generative-AI work. Uhmbrella's forthcoming tools will allow musicians to detect whether their content has been used in a particular piece of generative-AI music and even identify from which model the music was produced.[92] ProRata's attribution tools can take text and identify who the contributing authors are.[93] These tools will improve and others will become available as the industry matures, allowing rights holders who suspect that developers are abusing their results to test this and litigate accordingly.

Active poisoning tools such as Glaze and Nightshade[94] are exciting. The "poisoning" makes changes to the image that are invisible to the human eye but trick AI models into seeing cats as dogs and dogs as cats, for example. However, they are unlikely to work if universally adopted, because the underlying technology is too vulnerable to being reversed (and once reversed, the copyrighted content is then exposed to scraping until a new protection is rolled out).[95] While there is reasonable debate over the extent to which these defences might be saved,[96] these tools will probably remain a minor thorn for malicious developers who must take extra steps to ensure that the data they train on have not been poisoned.[97]

There will continue to be a market for defensive products but it is possible that there will be equity problems. Smaller rights holders and creative sectors may find that products are not available at a price and quality point that suits their needs. Given these products' status as tools for defending the public interest in the creative sector, governments may need to step in to fund the development of such tools if they are going to be viable at all. But it is very unlikely that they will be able to prop up an opt-out regime without strong standards and transparency.

07

# Establish Strong Standards for AI and Creativity

***Recommendation 3:*** *To safeguard the creative industries, clear standards must be established regarding creativity and licensing in AI applications. The UK government should introduce a one-off exception allowing major rights holders to license the past 75 years of content for AI training, as recommended in the AIOP. Additionally, standardised contracts between the creative and AI industries should be developed to enable more seamless business partnerships. The UK government must also provide clear guidelines on human creativity, ensuring that effective mechanisms exist to distinguish between human and AI-generated works.*

## Solving Licensing Problems

The UK has an enormous heritage of art and media that is not available on the open web. This work is worth billions but comes with a challenging rights problem. Each work will be tied up with dozens of rights holders, each exercising complex partial rights. It is almost impossible to work out whether the content owner has the right to relicense that work for AI training; as a result, much of this work goes unused.

For works being made today, contracts are already evolving to consider the downstream impacts of AI.[98] However, only governments can unlock archived content by granting a one-off exception to distribution for rights holders that allows them to relicense archived work for AI training without the explicit permission of all relevant rights holders. This has already been proposed in the AIOP.[99] This might seem like a significant giveaway to large content distributors, but this discussion transcends simple win-lose dynamics.

It is already difficult to establish licences for straightforward IP. The IP of a movie or a television series involves hundreds, sometimes thousands of other rights holders. If there is no one-off exception, this material will simply not be licensed and the opportunity for countries to influence training data sets with quality work from their culture will be lost forever.

In a similar way, downstream licences have many problems. While many media companies claim to want to license data widely, expectations around licensing terms vary widely. It is an essential part of all research (commercial and non-commercial) that the researcher does not know exactly how their data will be used in the final product. If they did, there would be nothing to research. However, rights holders sometimes insist that they are assured in advance of exactly how, where and when their data will be used. This shuts down many opportunities for licensing. This is particularly problematic for small rights holders on the web, for whom transactions costs will make it impossible to sign training licences without the adoption of automatic licensing technology.

Rights holders can be much more flexible in this landscape and accept that a good TDM licence is one with a considerable degree of uncertainty. Asking an end user to specify the model architecture in advance, or how many parameters the model will have, is not productive. Standardised contracts will also provide legal certainty about what can and can't be scraped and trained on, especially for the smallest rights holders.[100]

## Clarifying the Standards on Human Creativity

A key factor for the future of the creative economy will be the extent to which humans value art made by other humans over art created by machines. Salespeople have long recognised the value of the "handmade" label, despite the huge associated cost increases. Companies could celebrate the fact that they only contract human artists, in the same way as they pride themselves on the B Corporation label. Instances of companies using generative AI have gone down very poorly with customers.[101] It is not

the government's place to set this premium, but it is good economics to ensure that consumers have the information needed to distinguish one type of art from another.

The government must be able to set a consistent threshold for what qualifies as human, as opposed to being generated by computers. This issue is an old one: for example, digital sampling was initially considered an "unartistic" way of creating music because the musician did not directly create that sound. Fortunately, law already exists in most jurisdictions to distinguish the standard for human creativity. Artists are already exploring what it would take for AI-enabled art to receive copyright protections. For example, creative engine Invoke AI recently received the first successful copyright registration on what it called an "entirely" AI piece of art, entitled *A Single Piece of American Cheese*.[102]

However, the UK has a unique problem in this area because of its relationship with the EU. Currently there are two standards for human creativity that could be interpreted: the UK's "skill, judgement and effort" test and the "own intellectual creation" test, which derives from the European court. The merits of each test are beyond the scope of this paper, but the UK courts will have to contribute to establishing boundaries in the generative-AI era. The policy need is for legislators to review the law in this respect and ensure that consistent standards for human creativity exist that support the market for human-created work.

This would remove the relevance of the computer-generated works provision, which would also strengthen creators to a considerable extent and be consistent with other jurisdictions. This proposal is a small part of the Intellectual Property Office's consultation, but would prevent a situation where media giants, such as Spotify, can move from being a distributor to being a generative-AI producer. For example, any digital music that Spotify produces with generative AI would be exempt from copyright, and therefore Spotify would not legally be able to stop people making copies of its work. This disincentivises Spotify from displacing human artists entirely.

## Digital Watermarking

Consumers will not always be able to identify generative-AI products. Despite the presence of standards on human creativity, as discussed above, the final step in supporting the creative economy is to ensure that consumers can see what content is and isn't "human made". This is a problem for AI safety in general, especially in preventing the proliferation of deepfakes. This presents an opportunity for synergy between copyright and broader AI-safety policy.

Digital watermarks exist, the most well known of which is Google's SynthID.[103] Digital watermarks operate in a similar way to defensive tools such as Glaze and Nightshade. Changes to the image are made, invisible to the naked eye, that a machine can read as a sign that a piece of work is not human generated. This will allow users to identify and filter generative AI content on their browsers, social media and any other digital setting with the tools to read the watermarks.

There are two problems with the digital-watermarking approach. The first is that it is not clear whether it will work as well for text as it does for video and images. While Google claims that SynthID works well for text, OpenAI has been outspoken about the difficulties of detecting AI-generated work.[104] Second, it is hard to ensure that most generative-AI content is watermarked and that the watermarks stay on. While Google offers watermarking for content generated on its Vertex AI platform, there is nothing to prevent users on other platforms (or those running private models) from publishing content that is not watermarked. Many internet-hosting services can – and do – strip metadata such as watermarks from all content.

While perfect implementation remains the goal, consumers may find substantial value even with content watermarking that is 80 per cent accurate. Those that are keen to understand the provenance of their content are not likely to turn to X (which strips most metadata), while a business owner looking to buy art for their website may be pleased to know that checking content against a watermark engine can reveal which works have been made with the most popular AI tools.

The success of digital watermarking thus depends on the cooperation of the tech players who support a significant majority of internet traffic. If Microsoft, Google and Meta agreed that all generative-AI content being produced on their servers had to be watermarked, and remain watermarked while on their servers, this could enable consumers to identify the majority of generative-AI art that may be presented as having been created by a human. The video-games platform Steam already requires game studios to declare when they are using AI content.[105] OpenAI's new 4o image model labels all generated images with C2PA tags and can be identified by OpenAI internally as coming from 4o.[106] These moves ultimately support the work of artists by allowing them to market their work and enjoy whatever premium consumers place on human creativity.

Securing a thriving future for the UK creative industry is not dependent on predicting or controlling artistic innovation, but rather on establishing a fair market framework within which creativity can flourish in the generative-AI era. Whether artists ultimately incorporate AI as a complementary tool, maintain traditional approaches or develop proprietary models for licensing remains to be seen. What is essential for fair competition is preserving the distinction between intentional human creation and prompt-generated output.

08

# Support the Transition of the Creative Industries in the Generative-AI Era

***Recommendation 4:*** *The UK government should adopt a proactive approach to supporting the creative sector's transition into the AI era. This can be achieved through targeted funding and the establishment of a Centre for AI and Creative Industries (CACI). Given the complexities surrounding technological substitution and the legal implications of the Berne Convention – a treaty setting the framework for copyright law – a dedicated remuneration scheme should be introduced. Strategic investment in AI-ready data for the creative sector will also ensure that the UK remains at the forefront of AI-driven artistic and cultural innovation.*

## There Is Uncertainty Around the Impact of Generative AI on the Industry

A major challenge to designing policy around the issue of AI and copyright is uncertainty around future substitution effects. Generative AI may never be good enough to be a substitute for all human activities for which people get paid. For example, while Google's latest video engine Veo 2 can produce some lifelike clips, it is limited in the length of video it is able to produce, and still contains serious defects, especially around human hands and complicated mechanics. Suno, one of the most popular music-generating apps, ignores direct instructions about keys, time signature and scoring. The most sophisticated pieces of production software are tools for controlling art, rather than just generating it with a high degree of randomness.

Generative AI will continue to improve but its greatest value is likely to be in augmenting existing workflows, as demonstrated by platforms such as Invoke AI. If this is the case, many professional creators will keep their jobs, using generative AI as a novel source of material or to automate admin

tasks. Humans will also value content that is genuinely novel rather than incremental, and deep-learning models are not designed to provide this kind of innovation.

Lastly, while the debate on AI and copyright has been largely focused on the creative industries, generative AI that directly competes with such industries is unlikely to be a large part of the future AI economy. While the creative economy is about 5 per cent of UK GDP, every other part of the economy stands to benefit to varying degrees from AI diffusion. The narrative that developers are only the biggest US players and that all they want to do is replace artists is simply wrong; the positives of AI, both economic and scientific, will come from many more places.

## Establishing the Centre for AI and the Creative Industries

The UK boasts some of the world's most renowned educational institutions for the arts, including the Royal Academy of Arts, University of the Arts London and Glasgow School of Art, respectively ranked first, second and 13th globally in the 2024 QS World University Ranking for Art and Design.[107] Still, a critical skills gap exists. The 2024 Design and Copyright Society survey revealed that 96 per cent of artists have received no AI training, with 31 per cent citing this as a barrier to incorporating AI in their work.[108]

To bridge this gap, the UK government should establish a CACI. Arts schools, businesses, technology companies and policy centres currently operate in isolation; no dedicated hub exists where these disciplines can interact. This disconnect fuels concerns about AI's impact on creative professions and has hindered efforts to align technology policy with the creative industry in recent years.

The CACI would institutionalise the intersection of arts, industry, technology and policy. Led by international experts in creative AI, the centre would directly address current and future challenges facing the sector. By uniting leaders across these fields, it would drive innovation, enhance collaboration and solidify the UK's role in AI-driven cultural and economic development.

Support for such an initiative already exists. In October 2020, a consortium of academics and professionals, in partnership with the Nesta Policy and Evidence Centre, wrote an open letter to then-chancellor Rishi Sunak advocating for the centre. The letter, backed by economic and policy research from Nesta and the Creative Computing Institute (CCI), received positive feedback from the chancellor and leading academics, including the University of Edinburgh's Data-Driven Innovation cluster.

The centre would serve three functions: bringing together experts and representatives across what is becoming an increasingly entrenched and negative environment; acting as an engine to create new technologies and infrastructures to support growth in machine learning (ML) in the UK creative industries; and providing much-needed training and expertise across academia and industry. Similar centres at the intersection of computer science and the arts already exist in Europe, including IRCAM[109] and Ina GRM in France,[110] the Ars Electronica Center in Austria and ZKM in Germany. A UK-based centre with a specific focus on AI's role in the creative industries could foster collaborations with these European counterparts, while leveraging the UK's unique expertise.

Additionally, the CACI should have a remit for creating a national compute infrastructure to support creative industries in AI and ML, building and supporting new, transparently licensable models for use by artists, content creators and growing technology businesses. These models would have content licences supporting free or paid use. The CACI would use this resource to scaffold UK tech companies and creatives equally, while also serving as a best-practice example for AI's future role in culture. This could underpin a more positive and inclusive transformation of our technology and creative sectors.

## The Risk of Judicial Review

The UK, along with most countries, has signed the Berne Convention,[111] which requires – among other things – the principle of "national treatment", whereby works originating from one country have their copyright respected in another.

However, the convention limits governments' unilateral ability to create copyright exceptions. This comes from the three-step test (3ST), which says that governments cannot unilaterally grant copyright exceptions that interfere with the normal exploitation of a work. It is debatable whether a TDM exception without a remuneration system would pass the 3ST,[112] especially in the UK where a similar exception for private copying was dismissed after not including a private-copy levy that would have remunerated rights holders.[113] Similarly, empirical evidence that there are copyright regimes that do allow AI training and have not yet been hobbled by the 3ST suggests that the test is not a deal breaker for pro-AI copyright policies. However, it poses a risk proportionate to the extent that generative AI will affect the creative industries.

Though complete replacement of human creators appears unlikely, this ambiguity creates a degree of legal vulnerability for the UK government proposal, with the risk of a potential judicial review due to the absence of definitive evidence regarding AI's long-term effects on creative labour markets. Indeed, it seems likely that rights holders in the UK will pursue this pathway if the government goes ahead with policy suggested in the AIOP, as the Berne Convention also prohibits requiring any formalities for claiming copyright, of which the opt-out may be considered an example. A judicial review would not only delay the policy implementation and drain public funds through costly legal fees but could also completely invalidate the regulatory framework, forcing a comprehensive policy reconsideration and prolonging market uncertainty.

A related option would be to extend fair dealing. Certain copyright exceptions in the UK (such as private study) are limited by fair dealing, which among other things requires that copies do not infringe on the economic opportunities of the rights holder. This is very similar to the fourth of the criteria for fair use in the US and could be an option that enables non-competing AI technology to grow in the UK without affecting creators.[114] Crucially, developers could use content only if the final use of the model does not infringe on the economic interests of the rights holder. But it has

long been acknowledged that an extension of fair dealing to countries without a fair-use history could bring even more legal uncertainty and expensive litigation.[115]

The Berne Convention highlights the deep problem with any policy that relaxes copyright: with the best will in the world, the transition from the 20th-century copyright policy to one fit for 21st-century generative AI will be painful for many in the creative industries. It would be a reasonable political stance to smooth this journey with some support from the state, especially for a sector as socially important as the creative industry. This is not new thinking: Nobel Prize-winning economist Jean Tirole has long advocated for such policies.[116]

## The Benefits of a Remuneration Scheme

Introducing a targeted remuneration scheme would help address the question of remuneration, granting additional legal clarity in relation to the Berne Convention. This would be a significant win for developers and the government, while smoothing the economic transition for rights holders and addressing fairness concerns.

Assuming that any remuneration would not be taken out of existing government spending, money would have to be raised by a new or expanded mechanism, as covered earlier. Any such mechanism would have to be proportionate, independent, simple and efficient.

Any remuneration mechanism must not have unreasonable effects on the welfare of consumers, especially the poorest. From a distributive point of view, a widespread tax on consumer goods paid out to artists is at risk of being regressive, given that artistic professionals tend to come from advantaged socioeconomic backgrounds.[117] Policymakers should also avoid significant welfare impacts on consumers in general, as this would have an adverse effect on the economy and, especially, the technology ecosystem.

The mechanism should be independent of international partners. One challenge associated with governing AI and the internet is that these technologies are inherently global. With a new administration in the US positioning itself against AI regulation, the government must search for policy levers that do not require cooperation with unwilling partners or onerous legal challenges to technology multilaterals. Furthermore, different governments will have different priorities with respect to their creative and technological industries, so the mechanism must be adjustable to the needs of political leaders and their constituencies.

The mechanism must be easy for governments to operate and not create more of a bureaucratic burden. Complicated solutions that depend on arbitrary measures of sales or usage, for example, would be infeasible at a national level. The policy should also be independent of technology that may not yet exist. As discussed earlier, time is not on anyone's side, and the sooner an agreement can be reached the better.

The mechanism must also be efficient. It must not significantly distort the market for any services, and it should be especially sensitive to not making the UK a less competitive place for creators or technology developers.

## The Advantages of a Targeted Levy on ISPs

One option for funding a remuneration scheme would be to introduce a small levy on product that complements generative-AI consumption. This solution is not novel: so-called "private-copy levies" have existed for decades in Germany, have been investigated before as an option for the UK[118] and have been proposed by a consortium of CMOs.[119]

However, most private-copy levies work by taxing a tangible product, such as a device or storage medium. This creates a perverse incentive to buy these products from abroad, ideally avoiding any attempts to recoup the levy at the border. Furthermore, it is not the devices themselves that create and profit from the generative AI, so it is not clear why sellers of such devices should be bearing some of the tax incidence.

One alternative would be to tax AI vendors themselves. While this is desirable in the sense that it targets those who directly profit from generative AI, it has two distinct disadvantages. First, it will send negative signals about the government's growth vision for the AI industry. Even if the levy is minimal (less than 1 per cent), the media attention around such an announcement might deter some start-ups and investors. Taxing established industries is less likely to lead to such signalling problems, and defining AI will remain difficult. Second, AI is currently dominated by the US and China. Taxing tech giants from the US is likely to bring geopolitical pressures that the UK would want to avoid.[120]

A third option would be to tax data connections on fixed lines and mobile devices. These are non-fungible, creating no import pressure. An ISP levy is independent of other countries' policies on AI and copyright because it does not depend on governments' abilities to interact with technology companies. This would allow the UK government to avoid geopolitical controversies. The levy is also simple to implement: broadband services are easier to define than "smart devices" and are administered through a relatively small number of providers.

In addition, by targeting data connections, the levy does not encourage people to import devices from abroad; it is impossible to import broadband connections and importing mobile-data connections is too costly. The levy also avoids targeting technology companies directly, thus avoiding the criticism that it would reflect badly on the UK's tech investment.

An ISP levy comes with trade-offs, in that any tax on consumers is inherently anti-growth. However, the sums required for the ISP levy to be effective are very small – the equivalent to pennies per month for a regular household. Many consumers in countries with private-copy levies are unaware that the levies are ever included in their purchases. Furthermore, the levy is flexible to a country's particular financial situation: those that can afford to ask for more from consumers can do so, especially if they are supporting a larger creative sector.

Some may argue that the ISP levy is unfair because it assumes that consumers use generative AI equally. This is not the case, so a proportional levy would tax those who pay more for broadband and cellular-data connections. The question then is whether spending on these ISP services is associated with generative-AI consumption. While there are no publicly available data on this relationship, it is reasonable to assume that households and businesses that demand larger, faster data contracts are likely to be exposed to more generative AI, especially if one assumes that generative AI will proliferate widely over the internet over the coming years. Therefore, while there may be some consumers on ISP contracts who are not consuming generative AI, this number and the associated economic harm is likely to be small.

It might appear that the ISP levy has the wrong target. Given that the obvious beneficiaries of relaxed copyright legislation are AI developers, a levy on consumers seems backward. However, this ignores two important points.

First, AI developers develop because there is demand for their products. The final consumers of generative AI are consumers themselves. Therefore, it is not unreasonable to ask consumers (who are also stakeholders in the cultural life of the UK) to support the creative sector. Second, while there would be a hit to consumer welfare from an ISP levy, there would also be a small downstream effect on AI developers as people adjusted to the income effect of the ISP levy. Implementing an ISP levy is always going to be a political choice, with an impact on consumers and the wider economy, but the idea that the incidence of the tax is wrong is misguided.

## A Levy Would Raise Cover Funding for the Centre for AI and Creative Industries

The purpose of the levy would be to facilitate the transition of the creative industries into the generative-AI era in a socially progressive way, and to recognise the existence of bad actors in the scraping world. It would therefore not need to raise huge amounts to meet its goals.

Working on the basis that there are 116.1 million subscribers to mobile-data plans (including machine-to-machine) and 28.5 million broadband subscribers,[121] with monthly average subscriptions of £20 and £50 respectively, a tax rate of 0.1 per cent would yield total revenue of nearly £45 million. To reach a target revenue of £200 million, the tax rate could be increased to 0.44 per cent, resulting in consumers paying only about 31p extra per month.

Given the minimal size of the suggested levy, welfare effects on consumers would be marginal and could be reduced further by waiving the levy on broadband packages designed for low-income households (a similar policy was suggested for the Next Generation Fund).[122] As the tax is on consumers, there are no direct impacts on technology companies, apart from the reduction in revenue implied by the small income affect and intermediate effects on ISPs.

The priority of this revenue would be funding the CACI. This is an essential component in supporting the transition of the UK creative economy in the post-generative-AI world. This would still leave the UK government with hundreds of millions of pounds left over – about 10 per cent of England's central arts budget.[123] The remainder could be funnelled into existing central government bodies such as Arts Council England, with the caveat that such organisations have a long track record of disbursing grants to support the creative sector but have less to do with commercial art.

CMOs would be better suited to this task given their experience in distributing copyright royalites. They understand the structure of the sectors they represent and are better placed to make distributional decisions than central government. This is important when it is hard to answer questions about which rights holders are exploited the most by the AI-training process. All that the government needs to do is decide how to distribute the pot of levy funds to each CMO; how to distribute these new funds can be decided internally by the CMOs. Attribution tools such as those discussed above can be helpful here.

These problems have also been tackled before by countries implementing private-copy levies. Remuneration could also be directed to those artists who choose not to opt out, as a reward for their contribution to AI development. At a minimum, an infusion of funds in the order of millions of pounds per CMO is going to be a welcome bonus at a time of structural change in the creative sector.

The CACI could collaborate with legal representatives, content owners and tech providers to ensure fair compensation for AI-generated works. The centre would maintain transparent records of its data usage and analyse global AI models to assess how UK commercial data have been utilised, both with and without permission.

09

# Conclusion

The intersection of AI and copyright presents a profound challenge that will shape the future of the creative and technological industries. AI offers unprecedented opportunities for artistic expression and innovation, but also raises concerns among rights holders about consent, attribution and economic impact.

A restrictive copyright regime will not secure the future of the creative industries. Artists have historically responded very powerfully to technological advancements; from the advent of photography to digital sampling in music, creators have adapted and new forms of artistic innovation have flourished as a result. The same will hold true for the AI revolution, provided policymakers implement frameworks that enable rather than constrain creativity.

The implementation of opt-out regimes is a reasonable middle ground to meet the demands of rights holders and developers. Importantly, it would make it legal to train AI models on open-internet data while allowing rights holders the possibility to assert control over how their work is used. However, this policy option comes with some technical hurdles and implementation challenges. To overcome these, the UK government should work towards establishing clear standards for opt-outs, leveraging AI preferences protocols and promoting transparency among developers.

While AI and copyright disputes are often framed as a zero-sum game between rights holders and developers, this is a false dichotomy. A future for AI and creativity can be crafted that is not one of conflict but of collaboration. Rather than resisting AI, the creative community should be equipped with the skills and resources to harness its potential. Society is at the beginning of this journey, with AI already being integrated into creative workflows across industries, helping artists enhance their craft rather than rendering them obsolete.

The UK government must recognise that AI is a global technology and that restrictive copyright regimes in one country will not prevent the more open development of AI elsewhere. What it can do instead is act decisively to set up the UK creative and technological sectors for global success in the AI-driven economy, now and in the future. This means supporting AI-preferences standards, enabling flexible licensing agreements and investing in cutting-edge art training through initiatives such as the CACI.

The stakes are high. The UK's interpretation of copyright will play an important part in shaping the future of art, science and technological innovation. The government should embrace AI as an enabler of human creativity while implementing reasonable safeguards that ensure respect for rights holders. With these policies in place, the AI revolution can be the standout engine for artistic and cultural renewal of our era.

10

# Acknowledgements

The authors would like to thank the following experts for their input and feedback (while noting that contribution does not equal endorsement of points made in the paper).

Mat Dryhurst, Spawning AI

Lilian Edwards, Newcastle University

Tim Flagg, UKAI

Nate Hake, Travel Lemming

Meg Harding, PRS

Anton Howes, The Entrepreneurs Network

Eugene Huang, ProRata

Alys Key, Digital Frontier

Martin Kretschmer, CREATe

Chris Mammen, University College Oxford

Micaela Mantegna, Harvard metaLAB

Mark Nottingham, IETF

Kir Nuthi, Startup Coalition

Stephanie Reeves, British Copyright Council

Julia Rowan, PRS

Beatriz San Martin, Arnold & Porter

Megan Thomas, Meta

Lee Tiedrich, OECD

Philip Torr, Department of Engineering Science, University of Oxford

Ben White, Knowledge Rights 21

James Whittington, Thinking About Thinking, Inc

Julia Garayo Willemyns, UK Day One

*Lead image by Attilio Maranzano, courtesy of Fondazione Prada*

# Endnotes


1 https://theconversation.com/uk-arts-sector-is-getting-a-270-million-funding-boost-but-there-are-winners-and-losers-251340

2 https://avalon.law.yale.edu/18th%5Fcentury/anne%5F1710.asp

3 For a detailed description of training and the data collection lifecycle, see https://doi.org/10.52214/stlr.v26i1.13338

4 https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence

5 https://www.ft.com/content/a10866ec-130d-40a3-b62a-978f1202129e

6 https://rouse.com/insights/news/2024/artificial-intelligence-in-singapore-copyright-infringement-defence-for-artificial-intelligence-machine-learning

7 https://natlawreview.com/article/japans-new-draft-guidelines-ai-and-copyright-it-really-ok-train-ai-using-pirated

8 https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france

9 https://digitalcommons.wcl.american.edu/facsch%5Flawrev/1099/

10 https://www.reuters.com/technology/artificial-intelligence/deepseek-narrows-china-us-ai-gap-three-months-01ai-founder-lee-kai-fu-says-2025-03-25/

11 https://doi.org/10.52214/stlr.v26i1.13338

12 https://www.euronews.com/next/2025/01/24/openai-launches-first-ai-agent-operator-but-it-wont-be-coming-to-europe-yet

13 https://roost.tools/

14 https://www.beatoven.ai/blog/beatoven-ai-and-musical-ai-team-up/

15 https://www.kl3m.ai/

16 https://www.metmuseum.org/it/essays/the-daguerreian-age-in-france-1839-1855

17 https://www.fxguide.com/fxfeatured/inside-out-2-redefining-the-magic-with-new-technology/

18 https://www.vanityfair.com/hollywood/story/the-brutalists-ai-controversy-explained

19 https://www.gov.uk/guidance/the-rights-granted-by-copyright

20 https://semianalysis.com/2025/01/31/deepseek-debates/

21 https://www.eenewseurope.com/en/cerebras-launches-fastest-available-ai-inference-at-low-cost/

22 https://www.theguardian.com/artanddesign/2010/dec/01/tracey-emin-louise-bourgeois-

exhibition

23 https://www.theguardian.com/books/2013/jun/28/my-hero-louise-bourgeois-emin

24 https://www.fondazioneprada.org/project/serial-classic/?lang=en

25 While not a perfect compressor, the fact that model weights are a smaller (in terms of bytes) representation of their inputs means that they act somewhat like a compressor, like a .zip file of photos. See https://doi.org/10.48550/arXiv.1409.3215 and https://www.ibm.com/think/topics/variational-autoencoder

26 For example, if someone asks ChatGPT 4.5 for the first 1,000 words of ***The Handmaid's Tale***, the model refuses and cites copyright restrictions, even though copies of the book are available online.

27 https://www.bbc.co.uk/news/articles/c8xqv9g8442o

28 https://www.ons.gov.uk/economy/governmentpublicsectorandtaxes/researchanddevelopmentexpenditure/bulletins/ukgovernmentexpenditureonscienceengineeringandtechnology/2022

29 https://www.create.ac.uk/blog/2025/02/26/copyright-and-ai-response-by-the-create-centre-to-the-uk-governments-consultation/

30 https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth

31 https://ssir.org/articles/entry/ai-creativity-copyrights-patents

32 https://www.gov.uk/government/publications/pro-innovation-regulation-of-technologies-review-digital-technologies

33 We recognise that the TDM opt-out regime predates the AI Act in the Digital Single Market (DSM) Directive, but we refer to the AI Act as the most recent use of these ideas.

34 https://www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data%5Fd5241a23-en.html

35 https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change

36 https://publications.parliament.uk/pa/cm5901/cmselect/cmdfence/590/report.html

37 https://www.economist.com/briefing/2025/02/13/americas-military-supremacy-is-in-jeopardy

38 https://ollama.com/library/deepseek-r1:1.5b

39 https://roost.tools/

40 https://ed.newtonrex.com/optouts

41 https://institute.global/robots.txt

42 https://ed.newtonrex.com/robots.txt Note that the full page is much longer.

43 https://www.rfc-editor.org/rfc/rfc9309.html

44 https://aws.amazon.com/what-is/retrieval-augmented-generation/

45 https://laion.ai/

46 https://www.robotstxt.org/robotstxt.html

47 https://commoncrawl.org/blog/the-increase-of-common-crawl-citations-in-academic-research

48 https://www.turing.ac.uk/news/publications/understanding-artificial-intelligence-ethics-and-safety

49 https://www.nbcnews.com/tech/social-media/timeline-facebook-s-privacy-issues-its-responses-n859651

50 https://wordpress.org/plugins/dark-visitors/

51 https://datatracker.ietf.org/doc/rfc9309/

52 https://datatracker.ietf.org/wg/aipref/about/

53 https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

54 https://c2pa.org/

55 https://iscc.codes/

56 https://spawning.ai/browser-extension

57 https://github.com/Spawning-Inc/datadiligence

58 https://www.deviantart.com/team/journal/UPDATE-All-Deviations-Are-Opted-Out-of-AI-Datasets-934500371

59 https://doi.org/10.48550/arXiv.2404.02309

60 https://commoncrawl.org/

61 https://support.google.com/youtube/answer/2797370?hl=en

62 https://www.copyscape.com/

63 https://github.com/openai/openai-realtime-agents

64 https://doi.org/10.1162/qss%5Fa%5F00144

65 https://time.com/6247678/openai-chatgpt-kenya-workers/

66 https://doi.org/10.48550/arXiv.2005.14165

67 https://theodi.org/news-and-events/blog/ai-data-transparency-understanding-the-needs-and-current-state-of-play/

68 https://code-of-practice.ai/?section=transparency

69 For sub-URL content (such as particular paragraphs in text), the issue is even more complicated.

70 https://link.springer.com/chapter/10.1007/978-3-031-81596-6%5F14

71 This also explains why bulk licences might be difficult: the value of data will not be known until it has been tested as part of AI research.

72 https://github.com/google/robotstxt

73 https://doi.org/10.48550/arXiv.2012.07805

74 https://doi.org/10.48550/arXiv.2301.13188

75 Bing faces many of the same problems, but the focus here is on Google given its market share.

76 https://gs.statcounter.com/search-engine-market-share

77 https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers

78 https://finance.yahoo.com/news/chegg-sues-alphabet-saying-google-222229520.html

79 https://blog.google/around-the-globe/google-europe/an-update-on-googles-compliance-with-the-eu-copyright-directive/

80 https://apnews.com/article/google-gemini-ai-associated-press-ap-0b57bcf8c80dd406daa9ba916adacfaf

81 https://tollbit.com/bots/24q4/

82 The largest LLM with a publicly confirmed data set is Qwen2.5 with 18 trillion tokens. This is equivalent to nearly 169 million novels.

83 https://openreview.net/pdf?id=pEYxSx0frs

84 https://www.ietf.org/slides/slides-aicontrolws-control-starts-with-transparency-cloudflares-position-on-ai-crawlers-and-bots-00.pdf

85 https://world.org/

86 https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

87 https://www.venable.com/insights/publications/2022/08/are-your-website-terms-enforceable-maybe-not

88 https://lawcom.gov.uk/project/smart-contracts/

89 https://www.cloudflare.com/en-gb/application-services/products/bot-management/

90 https://kudurru.ai/

91 https://doi.org/10.48550/arXiv.2302.10149

92 https://musically.com/2024/11/27/uhmbrellas-ai-detection-tech-identifies-individual-platforms/

93 https://www.prorata.ai/

94 https://glaze.cs.uchicago.edu/index.html

95 https://nicholas.carlini.com/writing/2024/why-i-attack.html

96 https://glaze.cs.uchicago.edu/update21.html

97 The computational cost to "deglaze" images ranges from seconds to minutes, which is non-trivial for a crawler trying to clean nearly six billion images: https://laion.ai/blog/laion-5b/

98 https://www.equity.org.uk/advice-and-support/know-your-rights/ai-toolkit/equity-s-open-letter-to-the-industry-on-ai-training

99 https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

100 https://oecd.ai/en/wonk/standard-contract-terms-responsible-ai-data-sharing

101 https://www.pcgamer.com/wizards-of-the-coast-denies-using-ai-in-new-magic-the-gathering-image-this-art-was-created-by-humans/

102 https://www.invoke.com/post/invoke-receives-copyright-in-landmark-ruling-for-ai-assisted-artwork Not everyone would agree that this is "entirely" AI, because a human was involved throughout the process.

103 https://deepmind.google/technologies/synthid/

104 https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

105 https://store.steampowered.com/news/group/4145017/view/3862463747997849618

106 https://openai.com/index/introducing-4o-image-generation/

107 https://www.topuniversities.com/university-subject-rankings/art-design

108 https://cdn.dacs.org.uk/uploads/documents/News/Artificial-Intelligence-and-Artists-Work-DACS.pdf?v=1708424212

109 https://www.ircam.fr/

110 https://inagrm.com/en

111 https://www.wipo.int/treaties/en/ip/berne/

112 https://dx.doi.org/10.2139/ssrn.4629528; https://dx.doi.org/10.2139/ssrn.2331688

113 https://www.bailii.org/ew/cases/EWHC/Admin/2015/1723.html

114 Japan's "non-enjoyment" and "unjust harms" criteria also echo fair use/dealing. See https://natlawreview.com/article/japans-new-draft-guidelines-ai-and-copyright-it-really-ok-train-ai-using-pirated

115 https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth

116 https://press.princeton.edu/books/hardcover/9780691175164/economics-for-the-common-good

117 https://www.dacs.org.uk/news-events/earning-half-the-minimum-wage-new-report-reveals-pressures-on-artists-to-sustain-creative-life

118 https://www.gov.uk/government/publications/private-copying-and-fair-compensation

119 https://thesmartfund.co.uk/

120 https://thehill.com/policy/technology/5102654-trump-criticizes-eu-tech-fines/

121 https://www.ofcom.org.uk/phones-and-broadband/service-quality/communications-market-report-2024-interactive-data/

122 https://www.silicon.co.uk/e-regulation/digital-britain-6-levy-for-faster-broadband-1147

123 https://www.artscouncil.org.uk/our-organisation/annual-reports/arts-council-england-grant-aid-and-lottery-distribution-annual-report-and-accounts-202324

TONY BLAIR
INSTITUTE FOR
GLOBAL CHANGE

# Follow us

facebook.com/instituteglobal
x.com/instituteGC
instagram.com/institutegc

# General enquiries

info@institute.global