The Association of Photographers

AOP

AOP response to ICO Consultation Series on Generative AI and Data Protection: Chapter one - The lawful basis for web scraping to train generative AI models.

About the AOP

The Association of Photographers (AOP) exists to protect, promote, and inspire, championing the rights of all photographers and campaigning tirelessly on issues of copyright, best practice, and professionalism. Our 3,500 members include professional photographers, photographic assistants, photography agents, affiliated businesses, students, accredited photography courses at FE and HE level, and those working in support services for the Creative Industries. We are part of a greater network under the umbrella of the British Photographic Council which collectively represents around 15,000 creative professional image-makers in the UK.

The AOP membership has always been formed of some of the most influential, trailblazing photographers in the history of the art form. Past and present members include the likes of Terence Donovan, Rankin, Tim Flach, Nadav Kander, Tessa Traeger, David Bailey, Julia Fullerton-Batten and Jillian Edelstein. For over 50 years, members' work has appeared in global advertising campaigns, books, newspapers, magazines, exhibitions, and cultural events the world over.

Today, whilst our members explore and contribute to the development of the new realms of image technology at their disposal - the Association continues its mission to promote and protect the rights of individuals, which includes working closely with a range of All-Party Parliamentary Groups and creative industry representative organisations, such as the British Copyright Council (BCC) and Creators Rights Alliance (CRA), and importantly provides support to the next generations of photographers and image-makers through our close relationship with a growing number of universities and colleges.

Response to the ICO Consultation

Firstly, we very much welcome this timely ICO consultation reviewing generative AI and data protection, given the opacity behind which entities involved in data-scraping are undertaking such practices at significant scale and which often appear to be outside the UK's legal framework relating to GDPR, contract law and certain IP rights, specifically copyright.

Members of the AOP are mostly commissioned by Business-to-Business (B2B) sector clients which includes advertising and brands; however, our members also work with Business-to-Consumer (B2C) customers for print sales, books, and other merchandise, which means they handle personal data both in terms of the people they photograph and the B2C customers they deal with.

Our members are all encouraged to establish contractual terms of use for accessing and using services they offer through their websites, which includes machine-access that prohibits data-scraping without permission from the rightsholder. Their websites and platforms need to be open to the public (not to be confused with 'public domain') and cannot be placed behind logins, entry pages or paywalls, as no client or customer would entertain this approach to be able to view photographers' works. Without any form of contract override, or other technical protection measure, there is no technical way to limit lawful access without impacting normal business conduct.

We estimate the approximate number of images online displayed by UK professional photographers on their websites to be 362.76 million (avg.12,092, which may vary per photographer or image-maker, such as those with larger image archives accumulated over years), and the number of images licensed by professional photographers each year, to be approximately 114.76 million (AOP time-limited survey, conducted September 2022 and based on ONS stats 2022[1]). Most of these images are complete with metadata, which can identify individuals by name, geo-locations, and other personal details if a person is the subject in the image (made with their consent).

Web scraping

Since the arrival of generative AI software programs such as DALL-E, Midjourney, Stable Diffusion and others, our members have been deeply concerned about the extent to which their websites have been scraped for the rich data they contain, including personal data, copyright-protected works (containing metadata) and trademarks. In an address to a US Senate Judiciary Hearing on AI and copyright (July 2023), Stable Diffusion's Head of Public Policy, openly admitted to the amount and type of data they collect using the robot.txt protocol[2]. This protocol has existed for over 30 years as a type of 'gentlemen's agreement' in that there is no technical basis on which a web crawler or bot may be prevented from accessing website data. The terms of access that website and content owners put in place which may prohibit scraping for AI-purposes can be read by these crawlers and bots, but it is incumbent on those that send out the crawlers and bots to respect the terms of access. With AI-developers and data-miners seeking out an increasing amount of data, the trust in the use of this protocol is disintegrating. The key issue is that blocking access to web-crawlers is difficult to implement on an individual bot-by-bot basis and that blocking all simply impacts standard search engine optimisation (SEO), which is detrimental to those looking to promote themselves and their work online.

Lawful Basis

As far as UK copyright law is concerned, we firmly believe the acts undertaken by data miners and AI-developers in building their commercial interests are resulting in a significant level of copyright infringement of creators' works and breaches of data protection laws.

The framework for UK copyright legislation up to this point has often been heralded as a 'gold standard', being flexible enough to support innovation and drive the growth of incredible creative content to the extent that we are indeed significant global net exporters. Investment and economic growth in the UK creative industries is arguably a direct result of this framework, which has a long history in recognising and supporting our human creators. Even our framework

[1] ONS Statistics 2022 shows no. of photographers, audio-visual & broadcasting equipment operators is 73,300, therefore approx. 30,000 professional photographers is a fair assumption. https://www.statista.com/statistics/319286/number-of-photographers-audio-visual-and-broadcasting-equipment-operators-in-the-uk/

[2] US Senate Judiciary Committee Hearing on AI and Copyright [comments on Robot.txt 45:34-39: and 59:20-54] "We use robot.txt…digital standard which says I want my website to be used for ancillary purposes such as search engine indexing," https://www.judiciary.senate.gov/artificial-intelligence-and-intellectual-property_part-ii-copyright

for 'fair dealing' is a recognition of the skill, labour and judgement afforded to UK copyright works.

In addition, the UK's GDPR legislation affords vitally important protections for the handling of personal data, that our members adhere to, such as acquiring consent. However, because of data miners' and AI-developers' actions to ingest billions of images to train generative AI algorithms, without securing licensing permissions or obtaining consent for personal data processing, not only are the livelihoods of our members affected but also the people they photograph and provide services to.

We therefore strongly believe that data miners and AI-developers in the process of building their commercial interests are not complying with the lawfulness principle of data protection, which includes not breaching any laws, and having a valid lawful basis (Article 6(1) UK GDPR), which includes consent.

As for 'three-part' test demonstrating the following, we believe data miners and AI-developers building their commercial interests are failing in their lawful responsibilities. As highlighted below:

1. The purpose of the processing is legitimate;

A number of AI developers' approach to data scraping tends not to be framed for a specific purpose – whether they are explicit about what the model will be used for, and how any downstream use will respect data protection and people's rights, is unclear as the tech industry's default is to rapidly release products into the marketplace to outcompete rivals without robust safeguards and little formal regulation or oversight. AI-developers then claim their programs to be both for business interests and wider societal interests, but ignoring data subjects, creators and rightsholders protected by GDPR, and the harms caused by misrepresenting people either through information bias, deep fakes, copyright infringement and/or fraudulent activities. Large-scale and many smaller-scale generative-AI programs are developed with a lack of transparency, accountability and safety about how their data is obtained, what the purpose of a program is and what risk assessments and safety measures have been undertaken to reduce harm to users.

2. The processing is necessary for that purpose;

It is wholly inaccurate that most generative AI training is only possible using the volume of data obtained though large-scale scraping and there is evidence that generative AI could be developed with smaller, proprietary databases. For the image sector, we are already witnessing many smaller AI-developed text-to-image models made commercially available legitimately – namely, Getty Images[3], Shutterstock[4], Adobe[5], BRIA[6], Vaisual[7], for example, with the latter two companies having sought consent from contributors to the training datasets prior to training

[3] Getty Images Generative-AI https://www.gettyimages.co.uk/ai/generation/about;
[4] Shutterstock Generative-AI https://www.shutterstock.com/ai-image-generator;
[5] Adobe Firefly https://www.adobe.com/uk/products/firefly.html;
[6] Bria.ai (Fairly Trained approved) https://bria.ai/;
[7] Viasual.ai https://vaisual.com/;

their AI models. We also expect to see growth in smaller, curated and more accurate datasets using permission-based licensable images, as users (commercial and non-commercial) and consumers seek trusted sources that do not contain (and therefore do not replicate) information bias, deep fakes, copyright infringements and/or fraudulent behaviours.

3. The individual's interests do not override the interest being pursued.

We agree with the ICO's assessment of the 'process' by which data is collection is an invisible activity which most people are not aware of or have no control over the prevention of such activity. We have seen the increasing harms that social media has facilitated, affecting the use of people's personal data, which includes the use of images. Generative AI has the capability of scale and speed which only increases the risks and harm to people. Few safeguards are in place to prevent harms that can have a lasting impact on the person affected – whether these safeguards are couched in accountability, transparency, fairness or safety measures. We have already seen a massive upscale in deep-fakes of celebrities and historical/newsworthy moments that have been manipulated or misrepresented using generative-AI programs and may cause distress and reputational harm[8].

Risk mitigations to consider in the balancing test

There are few legal remedies and no obligations to make the owners of generative-AI platforms accountable as they tend to use the safe harbour principle/liability regime (present in US - s.512 of the DMCA, and EU law - E-commerce Directive 2000), a legal framework in which digital or online platforms are not legally responsible for hosting illegal content, but are required to remove such material once it is flagged. These liability regimes enable AI-developers to avoid their legal responsibilities to individuals affected by downstream harms. UK legislators could look to close this loophole giving regulatory powers to the ICO to be able to enforce accountability and ensure that technical and organisation guardrails and safety measures are in place and are effective for users and affected parties downstream. These safety measures and guardrails are presently missing.

Generative AI models deployed by the initial developer

Given our comments above about the lack of a UK legal framework to hold AI-developers accountable, who subsequently rely on other jurisdictions, and the examples given of those who are using smaller, licensed datasets that are GDPR-compliant and yet have not placed any

[8] A few examples - Taylor Swift deepfakes spark calls in Congress for new legislation https://www.bbc.co.uk/news/technology-68110476;
MrBeast and BBC stars used in deepfake scam videos https://www.bbc.co.uk/news/technology-6699365; Deepfake video of President Zelenskyy https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789; Artificially generated images of real-world news events https://www.washingtonpost.com/technology/2023/11/23/stock-photos-ai-images-controversy/; AI-Generated Influencer Emily Pellegrini deceives footballers and athletes https://then24.com/2024/01/03/this-is-emily-the-model-created-by-ai-who-has-deceived-elite-soccer-players-and-athletes/

apparent guardrails (footnotes), we do not believe or have faith that AI-developers can exercise complete control over how the generative AI model is used. We would support the introduction of regulatory powers which the ICO could leverage.

Generative AI models deployed by a third-party (not the initial developer), through an API

Where the initial generative AI developer can seek to ensure that the third party's deployment is in line with the legitimate interest identified at the generative AI training phase, again in our current experience, there is little control built in from the start. With DALL-E, Midjourney, and Stable Diffusion programs already on versions 3 to 6, and the machine-learning process already completed at the initial stage, there are no proven mechanisms to remove data (personal information or copyright-protected works) that has already been ingested and used to train these models.

In a research paper "Extracting Training Data from Diffusion Models"[9], published in January 2023, the researchers extracted over a thousand training examples from text-to-image diffusion models, ranging from photographs of individual people to trademarked company logos. The researchers' results found that diffusion models (used by DALL-E, Stable Diffusion and Midjourney) are much less private than previous generative models such as GANs (Generative Adversarial Networks), and that mitigating these vulnerabilities may require new advances in privacy-preserving training. They demonstrate that diffusion models do indeed memorise and regenerate individual training examples.

In the video recording of the US Senate Judiciary Hearing on AI and copyright (July 2023), Stable Diffusion's Head of Public Policy openly admitted that whilst there was a work-in-progress 'Opt-Out' process[10], they have not applied it to all programs, only new programs they develop. Older models retain the data, and they have not yet been able to remove any data [personal data and copyright-protected works] to these programs which are in service as it is not yet technically possible once a machine has been trained. This would support the concern that with text-to-image generative-AI programs, once they have been trained, programs cannot (yet) unlearn[11]. Machine unlearning applied to different data models is still an area of continued research. We would support further detailed research and assessment on whether future developments of Generative-AI models can specifically unlearn and what types of models these might be.

Generative AI models provided to third parties

For the reasons outlined in both the initial developer and third-party via API, we have little faith that accountability and safeguarding responsibilities are in place or will be implemented, without the pressure of regulatory legislative powers given to the ICO.

[9] Extracting Training Data from Diffusion Models, 30 Jan 2023 https://arxiv.org/abs/2301.13188

[10] US Senate Judiciary Committee Hearing on AI and Copyright [Parties comments on Opt-outs 01:19-1:25] https://www.judiciary.senate.gov/artificial-intelligence-and-intellectual-property_part-ii-copyright

[11] Now That Machines Can Learn, Can They Unlearn? https://www.wired.com/story/machines-can-learn-can-they-unlearn/

Additionally, we are deeply concerned about the use of “Open Source’ generative-AI models, without any regulatory accountability or safeguards. It is more than probable that the lack of control over personal data rights and intellectual property rights would be exacerbated significantly.

Conclusion

With the constant and regular scraping of personal information and copyright-protected works, without permission and in breach of legislation, and the speed and scale at which generative-AI can output material, we strongly believe that compliance with UK legislation on privacy, contract and IP law (copyright and trademarks) along with the introduction of additional regulatory powers that require transparency, accountability, safety and fairness, should be undertaken expediently.

The ICO was efficient and supportive in its delivery of the new GDPR regulations when they came into force in 2018, and we have confidence that the ICO will act to protect the rights of individuals in a fair and equitable way when it comes reviewing the lawful basis for web scraping.

Yours sincerely,

Isabelle Doran, CEO, The Association of Photographers
Nick Dunmur, Head of Business & Legal, The Association of Photographers

1 March 2024