ICO CONSULTATION RESPONSE

Response to Information Commissioner's Office – Draft Guidance on the Research, Archiving and Statistics (RAS) Provisions

Dr Zihao Li

Senior Lecturer in Technology Law; Zihao.Li@glasgow.ac.uk

Submitted based on the UKRI Metascience AI project: "Removing Legal Hurdles in Copyright and Data Privacy for AI-driven Research" (UKRI2603)

CREATe Subject Co-lead on Automation, Decentralisation and Platform

Date of submission: 5 May 2026

About the respondent

This response is submitted based on the UKRI Metascience AI project "Removing Legal Hurdles in Copyright and Data Privacy for AI-driven Research: Unleashing the Potential of AI for Science", led by Dr Zihao Li (Senior Lecturer in Technology Law, University of Glasgow), with affiliation to the Centre for Regulation of the Creative Economy (CREATe), an interdisciplinary research centre anchored in intellectual property, competition, information and technology law. CREATe is funded as AHRC research infrastructure.

The UKRI project investigates how the UK's copyright and data protection frameworks shape AI-driven research across the full research lifecycle, combining empirical investigation (a national survey of UK researchers and qualitative interviews), techno-legal analysis of the CDPA, the UK GDPR and the DUAA 2025, and comparative analysis with EU and US frameworks. The project will produce policy recommendations for UKIPO, the ICO and UKRI. The project started in October 2025. This response addresses the data protection issues and will offer further empirical insights as the project progresses. We would welcome the opportunity to engage further with ICO officials, including through workshops and the project's planned stakeholder events.

1. Executive summary

This response is submitted on behalf of the UKRI Metascience project "Removing Legal Hurdles in Copyright and Data Privacy for AI-driven Research: Unleashing the Potential of AI for Science", hosted at the CREATe Centre and the School of Law, University of Glasgow. The project examines how the UK Copyright, Designs and Patents Act 1988 (CDPA), the UK GDPR and the Data Protection Act 2018 (DPA), as amended by the Data (Use and Access) Act 2025 (DUAA), structure the legal conditions for AI-driven scientific research across the full research lifecycle (acquisition, organisation, training, deployment, verification, publishing, reuse).

We welcome the ICO's draft guidance on the Research, Archiving and Statistics (RAS) provisions. The guidance offers valuable clarification on a number of points that have hampered AI-driven research, including the recognition that scientific research extends to commercial activity, the explicit treatment of broad consent, the compatibility of further processing under Article 8A UK GDPR, and the affirmation that consent under Article 6(1)(a) is, in most cases, not the appropriate lawful basis for research processing. These positions are consistent with the leading academic literature on the data protection treatment of scientific research and they directly address several longstanding sources of risk-aversion in the UK research community.

At the same time, the draft guidance, taken as a whole, continues to map data protection obligations onto a static, project-by-project, individualised model of research. That model does not describe modern AI-driven research, where datasets are shared, reused, model weights are derived, retraining is iterative, and pipelines often span universities, cultural institutions and industry partners. Several elements of the draft, in particular the four criterion test for "scientific research" (scientific objective, scientific method, uncertainty, transferability), the strict reading of the Article 84C exclusion of "measures or decisions about particular people", and the case-by-case approach to exemptions from data subject rights, risk tightening rather than easing the eligibility threshold for legitimate AI research.

Our submission therefore makes eight principal recommendations, summarised here and developed in section 4:

- Adopt a lifecycle-aware reading of "scientific research" that recognises sharing, training, validation, retraining and benchmarking as integral acts of research, not merely scientific research, engineering or product development.
- Treat the four criteria (scientific objective, method, uncertainty, transferability) as cumulative indicators rather than gates; apply them flexibly to applied AI research, including industry academic collaboration.
- Confirm explicitly that pre-deployment AI development (including model training, fine-turning, distillation), including by commercial entities, falls within Article 4(2) UK GDPR "scientific research", with heightened obligations crystallising at market entry rather than at the point of training.
- Clarify the interaction between the Article 84C "measures or decisions about particular people" safeguard and AI research that may, downstream, inform automated decision-making, so that the safeguard does not chill upstream research that has no operational decisional purpose.
- Provide concrete guidance on the application of the disproportionate effort exception in Article 14 to large scale and web scraped corpora used in AI training, including indicative factors and worked examples.
- Recognise that web scraped, aggregated and inferential data raise distinct lawful basis and special category issues that the current draft only partially addresses.
- Coordinate the RAS guidance with the parallel reform of the CDPA s.29A text and data mining (TDM) exception, so that copyright law and data protection law impose consistent, not contradictory, obligations along the AI research lifecycle.

- Develop a "research safe harbour" phase: lenient obligations inside the research lifecycle (subject to internal governance, retention, security and oversight), and concentrate intensified obligations (transparency, accountability, downstream redress) at the point of market entry.

These recommendations are evidence-based, drawing on the project's ongoing empirical and techno-legal analysis (Martin et al., 2025; CREATe submissions to UKIPO consultations, 2022 and 2025), as well as the wider academic literature on data protection and scientific research (Erdos, 2011; Ducato, 2020; Malgieri, 2020; Meszaros & Ho, 2021; Solove, 2024).

2. General observations on the draft guidance

2.1 What the draft guidance does well

We note and welcome four positive features of the draft guidance that materially advance the legal conditions for AI-driven scientific research in the UK:

- **Recognition of commercial scientific research.** The reaffirmation, drawing on Article 4(2) UK GDPR, that scientific research "whether publicly or privately funded and whether carried out as a commercial or non-commercial activity" falls within scope is a vital and welcome alignment with the recent AI development. This corrects a perception, persistent in research ethics practice, that scientific research is essentially synonymous with publicly funded academic research.
- **Clarity on lawful basis.** The explicit statement that consent under Article 6(1)(a) is, in most cases, not the appropriate lawful basis for research related processing, and that public task or legitimate interests will normally be more appropriate, is consistent with both the prevailing academic consensus (Meszaros & Ho, 2021; Ducato, 2020) and the practical reality of large scale AI research. This will help dispel a longstanding source of uncertainty in the UK research community.
- **Article 8A compatibility presumption.** The position that further processing for RAS purposes is presumptively compatible with the original purpose, where the safeguards in Article 84B/84C are met, removes a significant source of uncertainty for projects that wish to reuse existing institutional or partner data for AI research.
- **Broad consent and storage limitation.** The guidance highlights that "broad consent" is permissible where the specific research purposes cannot be fully identified ex ante, and the indefinite storage permitted under Article 5(1)(e) for RAS purposes (subject to safeguards), are pragmatically important for iterative AI research and longitudinal datasets.

2.2 Where the draft guidance still insufficient

Despite these welcome elements, the draft as a whole continues to assume an individualised, project-bounded model of research that does not match the realities of AI-driven research. The literature on data protection and scientific research has long identified the same problem in the underlying GDPR architecture (Ducato, 2020; Erdos, 2011; Malgieri, 2020), and the DUAA reforms, helpful as they are, only partially address it. As a result, four structural difficulties persist in the draft guidance:

First, the four criterion test for scientific research, *as currently framed,* **tightens rather than eases** the eligibility threshold for AI research. The criteria of "scientific objective" (improvement in science or technology that benefits the entire field), "uncertainty" (the outcome is not obvious in advance) and "transferability" (new knowledge that can be shared) are individually defensible but collectively risk excluding important categories of legitimate AI research, including incremental model improvements, model fine-turning, model distillation, benchmarking, replication studies and applied research carried out within companies. We address this in detail in section 4.1.

Second, the strict reading of "measures or decisions about particular people" in Article 84C, while doctrinally commendable, is liable to be applied in a way that chills upstream AI research that does not, of itself, take operational decisions about individuals, even when the eventual deployed model might. The current draft does not adequately distinguish research stage processing (which is exploratory and aggregate) from deployment stage processing (which may be decisional). Especially for deployment stage processing, we might need stricter mechanisms.

Third, the case by case requirement for exemptions is, in principle, the correct approach. In practice, however, when combined with overly conservative DPIA cultures (Erdos, 2011; Ducato, 2020), it produces precisely the risk-averse, default refusal behaviour that the RAS provisions are intended to dispel. The guidance should support compliance by design workflows for AI research, in which standard institutional safeguards discharge the substantive requirements of Article 84B/84C without requiring per data subject reasoning.

Fourth, the guidance is silent on the interaction with copyright law, in particular the CDPA s.29A text and data mining exception, the sui generis database right, contractual overrides and technological protection measures. AI research is shaped by both regimes simultaneously. Guidance that treats data protection in isolation will leave researchers caught between two non-aligned frameworks, both of which can independently chill the same research activity.

3. The Metascience and AI-for-Research lens

Our project starts from a metascientific premise: AI is not only an object of research; it is, increasingly, the means by which research is conducted across all disciplines, from computational biology, climate modelling to digital humanities and social science. This shift, which we describe as "AI for Research" (AI4Research), reframes how data protection law applies to scientific research.

Three features of AI-driven research are particularly relevant to the RAS guidance:

- **Lifecycle shaped processing.** AI research follows a multi-stage lifecycle (acquisition, organisation, training, deployment, verification, publishing, sharing, reuse). The same statute and the same data may engage data protection law differently at different stages. Guidance that treats research as a single, monolithic activity will produce blunt outcomes.
- **Mixed and inferential data.** AI training corpora often combine personal, non-personal, public, scraped, synthetic and inferential data. Special category and criminal offence data

may be present in non-obvious ways (e.g., revealed by inference from non-sensitive features). Provisions on special category data must remain workable in this setting.

- **Distributed actors and supply chains.** AI research is rarely a single controller project. Universities, cultural institutions, infrastructure providers, AI agents and industry partners share data and models. The controller / processor / joint-controller architecture of UK GDPR is destabilised by these supply chains, and the guidance should provide clarity on how the RAS provisions operate in collaborative settings.

These features are not edge cases. They describe the modal form of contemporary scientific research that uses AI, and the empirical evidence from literature shows that the failure of the law to engage with them produces fragmented datasets, abandoned collaborations, and systematic blind spots in research on issues such as discrimination, extremism and structural inequality (Erdos, 2011; Ducato, 2020).

4. Detailed comments and recommendations

4.1 The four criteria for scientific research

We welcome the inclusion of fundamental research, applied research and experimental development. However, we have three concerns about the four criterion test.

(a) The "scientific objective" criterion is too demanding for applied AI research and social and humanity research

The draft requires that the work "aim to make a meaningful improvement in science or technology" that "benefit[s] the entire field, not just the organisation doing the research", and the exclusionary evidence includes "the work is simply copying, tweaking, or reverse engineering an existing technology" and "a skilled professional could easily achieve the improvement being pursued using publicly available information."

In AI research, much legitimate scientific work is, on its face, incremental: fine-tuning a base model on domain data, distilling a model for downstream use, replicating a published result on a new corpus or context, benchmarking, ablation studies, or building reference datasets. The empirical literature in machine learning explicitly recognises these as scientific contributions. A criterion that excludes "copying, tweaking, or reverse engineering" risks excluding much applied AI research and reproducibility work. Moreover, it should also more fully recognise the contribution of social science and humanities research, where originality often lies not in building a novel model, but in evaluating impacts, interrogating assumptions, contextualising technologies, and generating critical or interpretive knowledge about their operation and effects, especially in some extreme studies (e.g., hate speech, discrimination studies). Similarly, some forms of research, such as work on traditional medicine, may struggle to be recognised as "scientific research" under a narrow or overly technocentric understanding of that term.

Recommendation 4.1(a). The guidance should clarify that incremental contributions, including fine-tuning, transfer learning, replication and benchmarking, can satisfy the "scientific objective" criterion

where they are carried out with scientific rigour and produce knowledge that is in principle transferable. The exclusionary evidence should be reframed as a presumption rebuttable by methodological evidence, not as a hard exclusion. Moreover, it should broaden the scope of "scientific" or consider using the more broad term "research" to replace.

(b) The "uncertainty" criterion misreads contemporary AI research

The draft excludes work where "the only uncertainty is commercial", where "the work merely customises existing technology, unless it's genuinely unclear whether the customisation is even possible", and where "the work involves building a prototype, unless there is real uncertainty about how to do it."

AI research is often characterised by empirical uncertainty rather than algorithmic uncertainty. The fundamental uncertainty in many machine learning experiments lies in whether a particular architecture, dataset or training regime will generalise to a new setting. That is genuine scientific uncertainty, even if the architecture is itself well known. The exclusion of "mere customisation" as a default risks misclassifying these experiments as engineering.

Recommendation 4.1(b). The guidance should expressly recognise empirical uncertainty about generalisation, transfer, robustness and performance on AI research as forms of scientific uncertainty, even where the underlying architecture is known. The exclusion for "customisation" should not be applied to AI research that uses customisation as the experimental object of inquiry.

(c) Treating the four criteria cumulatively risks narrowing the gateway

The draft does not specify whether the four criteria operate cumulatively (all must be met) or as indicators (the totality of the evidence supports the characterisation). If applied cumulatively, the test will narrow eligibility considerably, particularly for industry–academic collaboration and for the increasingly common pattern of agentic AI research where data and methods are selected dynamically at run time.

Recommendation 4.1(d). The guidance should clarify that the four criteria are indicative and weighted in the round, not cumulative gates.

4.2 Commercial vs. academic research

We strongly welcome the affirmation that scientific research is not limited to non-commercial activity. This aligns the data protection treatment of research with both the EU GDPR architecture and the realities of contemporary research, where universities, infrastructure providers and industry partners frequently co-produce data, models and findings (Meszaros & Ho, 2021).

However, this position sits in tension with the parallel UK copyright regime, where the s.29A CDPA TDM exception is restricted to non-commercial purposes. Researchers in public-private partnerships are therefore in the paradoxical position of qualifying as scientific researchers under data protection law while losing the benefit of the copyright exception that would otherwise authorise the underlying acts of reproduction. Our project develops this point in detail.

Although this is not a matter of data protection guidance alone, it has direct operational consequences for the ICO's guidance. Researchers who cannot lawfully make a copy of training data under copyright law will be unable to rely on the Article 8A compatibility presumption or the Article 84B requirements, no matter how clearly drafted the RAS guidance becomes. The two regimes must be read together.

Recommendation 4.2(a). The ICO should engage with the UKIPO and the Department for Science, Innovation and Technology to ensure that the s.29A CDPA reform process and the RAS guidance produce coherent, lifecycle aligned obligations. The two frameworks should not drift further apart.

Recommendation 4.2(b). The RAS guidance should include a worked example covering a public-private AI research collaboration that scrapes copyrighted material and personal data from the open web, addressing how the relevant GDPR provisions apply across the different parties. We are happy to provide such an example based on our empirical work.

Recommendation 4.2(c). The guidance should make clear, in line with Article 4(2) UK GDPR, that pre-deployment AI research conducted by commercial entities (including foundation model training and red teaming) is capable of falling within "scientific research", with heightened transparency, licensing and accountability obligations crystallising at the point of market entry rather than at the point of training. This mirrors the "research safe harbour" concept set out in section 4.7.

4.3 Lawful basis: legitimate interests, public task and the legitimate interests assessment

We welcome the position that legitimate interests and public task will normally be the appropriate bases, and that consent is generally inappropriate. We also welcome the statement that, where the processing has appropriate safeguards, ethical approval and regulatory compliance, organisations "can generally be confident that the legitimate interests lawful basis applies" and need not undertake a separate LIA.

This is a pragmatic and important position. We recommend that it be developed slightly further to accommodate AI research.

Recommendation 4.3(a). The guidance should expressly state that compliance with established institutional safeguards (DPIA, ethical approval, trusted research environments, ISO/IEC relevant standards, cybersecurity assessment) is sufficient to discharge the substantive concerns that an LIA would address, and that no separate LIA is required where these safeguards are demonstrably in place. This will reduce duplicative paperwork that, in practice, drives risk-aversion.

Recommendation 4.3(b). The guidance should clarify the application of the recognised legitimate interests gateway under Article 6(1)(ea) introduced by the DUAA 2025 to AI research. In particular, the relationship between the recognised legitimate interests list, the research related processing test and the safeguards under Article 84B/84C should be made explicit, including for cases where the same processing might satisfy more than one gateway.

Recommendation 4.3(c). Where research uses publicly accessible data scraped from the open web, the guidance should engage directly with the lawful basis question. The literature is well aware that

public availability does not equate to lawful basis (Kuru, 2024; Solove, 2024). We recommend that the ICO's recent generative AI consultation outcomes be cross referenced and applied to the research context, including for scraped corpora used in academic and industry-led model training.

4.4 Article 84C safeguards – substantial damage, decisions about people, and AI research

We support the substantive safeguards in Article 84C: prohibition of substantial damage or substantial distress, prohibition of measures or decisions about particular people, and the data minimisation requirement. These reflect long standing principles in the academic literature (Malgieri, 2020; Ducato, 2020). However, two clarifications are essential for AI research.

(a) The "measures or decisions about particular people" limit

The draft states that the RAS provisions cannot be used "if you are carrying out processing for the purposes of measures or decisions with respect to the people whose information you are using, unless the research is approved medical research", and adds that an organisation cannot reuse RAS-retained data for downstream operational decisions.

This is, in our view, correctly stated as a matter of doctrine. However, it requires careful application to AI research, where the same upstream training process may, in due course, give rise to a deployed model that does take decisions about individuals. We agree that the deployed model's use is not research, but the upstream training, evaluation, validation and fine-tuning should not be treated as a "measure or decision" about the training data subjects merely because the eventual product is decisional.

Recommendation 4.4(a). The guidance should clarify that processing during the research lifecycle (training, evaluation, validation, retraining) does not constitute "measures or decisions about particular people", even where the eventual deployed model may, post market entry, inform such decisions. The RAS provisions should remain available upstream, and the relevant safeguards (transparency, contestability, automated decision making protections under Article 22 UK GDPR) should attach at deployment, not at training. Moreover, it should protect individuals' collective interest rather than treating them isolated.

(b) Substantial damage and substantial distress in AI research

Defining "substantial damage" as including "loss of confidentiality" and "deprivation of rights" is appropriate but, applied without nuance, could be read to capture any AI research on web scraped corpora. The literature (Erdos, 2011; Ducato, 2020) shows how broad readings of risk of distress have produced systematic blind spots in research on extremism, discrimination and politically sensitive issues.

Recommendation 4.4(c). The guidance should expressly recognise that contemporary historically and politically sensitive research (extremism, hate speech, structural discrimination, political radicalisation) cannot be conducted without engaging with sensitive material, and that the substantial damage /

substantial distress test must not be applied so as to render such research impossible. A worked example would be helpful.

4.5 Exemptions from data subject rights

We welcome the clarification on this matter. We are concerned, however, that the practical effect of this requirement, in combination with under-resourced data protection offices in many research organisations, will be the entrenchment of risk-averse default refusal behaviours.

(a) Right to be informed and the disproportionate effort exception

Article 14 contains a critical valve for AI research: where personal information has been collected from a source other than the data subject, the obligation to inform may be lifted where doing so would involve "disproportionate effort" or would "likely make impossible or seriously impair the achievement of the processing's objectives". The draft guidance gestures at this but does not, in our view, give it the operational specificity that AI research requires.

Web scraped training corpora often contain millions of data subjects across many jurisdictions; individual notification is often impossible in any meaningful sense. The literature recognises that the GDPR's individualised transparency model does not scale to such corpora (Ducato, 2020; Solove, 2024; Kuru, 2024). The DUAA 2025 has clarified the disproportionate effort exception, and the guidance should fully exploit this clarification.

Recommendation 4.5(a). The guidance should set out indicative factors and worked examples for the application of the disproportionate effort exception in the context of large scale AI training corpora, including (i) corpus-level transparency through prominent project webpages, opt-out portals, and dataset cards; (ii) the role of intermediaries (publishers, platforms) in conveying privacy information; (iii) emphasising the function of pseudonymisation, deletion on request mechanisms and trusted research environments, which should be regarded as a compliance and protection measures.

(b) Right of access, rectification, restriction, erasure and objection

The exemptions in DPA 2018 schedule 2 paragraphs 27 and 28, and the built-in exception in Article 17(3)(d) UK GDPR, cover most foreseeable AI-research scenarios where compliance with the right would seriously impair the research. We support the case by case requirement in principle, but recommend operational guidance:

Recommendation 4.5(b). Where research is carried out within a trusted research environment, with documented governance and a published privacy notice, the guidance should recognise that the case by case assessment can be discharged through an existing standing risk assessment framework. Researchers should not be required to repeat the same reasoning for each individual request where a documented institutional procedure already exists.

Recommendation 4.5(c). The guidance should explicitly recognise that, for trained model weights, the right to erasure cannot in many cases be operationalised without retraining the model from scratch – a step that would seriously impair the research and is sometimes computationally infeasible. The

Article 17(3)(d) exception should be read to cover this scenario, with corresponding obligations focused on the upstream training corpus and downstream output controls (e.g., output filtering, watermarking, machine unlearning research).

4.6 Special category and criminal offence data in AI research

AI training corpora frequently contain special category data, sometimes inadvertently, and inferential outputs can reveal special category information from non-sensitive inputs (Solove, 2024; Bekkum & Borgesius, 2023). Article 9(2)(j) read with schedule 1 paragraph 4 DPA 2018 provides the necessary research condition, but its application to AI corpora needs clarification.

Recommendation 4.6(a). The guidance should clarify that incidental special category data in a general purpose research corpus may be processed under Article 9(2)(j) where (i) the corpus is processed for genuine research purposes, (ii) the special category data is not the principal target of the processing, and (iii) appropriate technical and organisational safeguards are in place. The alternative, requiring exhaustive ex-ante audit of web scraped corpus, is not feasible and would chill all such research.

Recommendation 4.6(b). The guidance should engage with the question of inferential special category data: when, if at all, the model's ability to infer special category attributes makes the inputs themselves special category data. This is a live question and one of the most difficult issues for AI researchers.

4.7 A research safe harbour: lenient inside, account at the edge

Building on the project's techno-legal analysis, we propose that the ICO articulate, alongside the case by case approach and working with UK IPO, a structural "research safe harbour" concept that integrates the data protection and copyright treatment of AI research.

The safe harbour concept has two components:

- **Inside the lifecycle (acquisition through verification):** a unified research exemption across the CDPA s.29A TDM exception and the UK GDPR RAS provisions, conditional on internal governance (purpose limitation, retention rules, security, oversight, ethical approval, controlled access). Within these conditions, researchers can collect, retain, share with collaborators, train and evaluate, with reduced transparency and rights-based obligations to data subjects.
- **At the edge (market entry):** heightened obligations crystallise. Transparency, licensing or equitable remuneration, accountability for downstream harms, automated decision-making safeguards, and full data subject rights apply at the point of first external deployment, paid access or downstream commercial licensing.

This concept can be implemented within the current statutory framework, by aligning the ICO's reading of "scientific research", the Article 84B/84C safeguards, and the disproportionate effort exception, with the IPO's reform of s.29A CDPA. It does not require primary legislation, but it would benefit from explicit endorsement by the ICO and IPO.

Recommendation 4.7. The ICO should adopt, in its final guidance, a lifecycle-aware reading of scientific research that distinguishes inside the lifecycle research processing from at the edge of market entry processing, and align the two with corresponding obligations. We are happy to provide further detail through the project's ongoing techno-legal analysis.

4.8 Coordination with other regimes

Finally, we observe that the AI research lifecycle is regulated by several overlapping regimes: copyright law (CDPA s.29A; sui generis database right; contractual overrides; technological protection measures), data protection law (UK GDPR, as amended by DUAA), the Equality Act 2010, sector specific regimes (e.g., medical regulations, NHS policies), and the emerging AI-specific governance landscape (the AI Opportunities Action Plan, the EU AI Act through its extraterritorial reach).

Guidance that treats the data protection dimension in isolation is, by construction, only part of the picture. We urge the ICO to coordinate with UKIPO, UKRI, and the Digital Regulation Cooperation Forum (DRCF) to ensure that the operative obligations on researchers across these regimes are consistent.

Recommendation 4.8. The final guidance should reference, where relevant, the parallel reform of relevant regimes, such as copyright law, healthcare regulation, and the UKRI and other funder's policy.

5. Conclusion

The ICO's draft guidance is a substantial and welcome step. It clarifies several long standing sources of uncertainty, including the commercial scope of "scientific research", the appropriate lawful basis, broad consent, and the storage limitation position. These clarifications will materially improve the legal conditions for AI-driven research in the UK.

The principal remaining issue is structural. The data protection treatment of research, even as updated by the DUAA, continues to be modelled on individualised, project bounded research. AI-driven research is iterative, multi-stage, multi-actor and large-scale. A guidance document that does not engage with this structural shift will, in practice, leave researchers caught between an over-regulatory upstream and an under-protective downstream, the diagnosis we describe in our paper as "too restrictive, too permissive".

We urge the ICO to use the final guidance to articulate a lifecycle-aware reading of scientific research, to coordinate with the parallel copyright reform, and to recognise the legitimacy of compliance by design workflows for AI research. Done well, the guidance can position the UK as the leading jurisdiction for responsible, innovative AI-driven research, advancing both the UK Government's AI Opportunities ambitions and the protection of data subjects.

We would welcome the opportunity to engage further with ICO officials on any of these points, and to share interim findings from the project's empirical investigation.

6. Selected references

Bekkum, M. & Borgesius, F. Z. (2023). Using sensitive data to prevent discrimination by artificial intelligence: does the GDPR need a new exception? Computer Law & Security Review, 48, 105770.

CREATe (2025). Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation. CREATe, University of Glasgow.

Ducato, R. (2020). Data protection, scientific research, and the role of information. Computer Law & Security Review, 37, 105412.

Erdos, D. (2011). Stuck in the thicket: social research under the first data protection principle. International Journal of Law and Information Technology, 19(2), 133–152.

Kretschmer, M., Margoni, T. & Oruc, P. (2024). Copyright law and the lifecycle of machine learning models. IIC - International Review of Intellectual Property and Competition Law, 55, 110-138.

Kretschmer, M. et al. (2025). Copyright and AI in the UK: Opting-In or Opting-Out? GRUR International, 74(11). 1055-1079.

Kretschmer, M., Meletti, B. & Porangaba, L. H. (2022). Artificial intelligence and intellectual property: copyright and patents – a response by the CREATe Centre to the UK Intellectual Property Office's open consultation. Journal of Intellectual Property Law & Practice, 17(3), 321-326.

Kuru, T. (2024). Lawfulness of the mass processing of publicly accessible online data to train large language models. International Data Privacy Law, 14(4), 326-351.

Malgieri, G. (2020). Data protection and research: a vital challenge in the era of COVID-19 pandemic. Computer Law & Security Review, 37, 105431.

Margoni, T. & Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR International, 71(8), 685–701.

Meszaros, J. & Ho, C. (2021). AI research and data protection: can the same rules apply for commercial and academic research under the GDPR? Computer Law & Security Review, 41, 105532.

Quintais, J. P. (2025). Generative AI, copyright and the AI Act. Computer Law & Security Review, 56, 106107.

Solove, D. J. (2024). Artificial intelligence and privacy. Florida Law Review, 77(1).

Lin, H. & Zhang, Y. (2025). Navigating the Risks of Using Large Language Models for Text Annotation in Social Science Research. Social Science Computer Review.