BIA policy submission

25 February 2025

UK BioIndustry Association

BIA submission: Copyright & AI

About the BIA and our members

The BIA is the voice of the UK's innovative life science and biotech industry, and our mission is to enable and connect the UK ecosystem so that businesses can start, grow, and deliver world-changing innovation. The BIA has a diverse membership, counting over 600 members including start-ups, scale-ups and established global companies, as well as universities, research centres, and investors.

Life science is a growing sector of the future that poses a unique opportunity. The UK life sciences industry employs over 300,000 people, with around two-thirds of these jobs outside London and the South East. There are 6,850 life sciences businesses, 75% of which are SMEs, and combined they generate a turnover of £108.1bn.[1] The average GVA per employee is over twice the UK average at £104,000 and the sector consistently invests more in R&D than any other (£9 billion in 2022).[2]

BIA primarily represents innovative start-ups and scale-ups, including those that use AI and data in drug discovery, medicines development, and diagnostics, which is a rapidly growing field referred to as 'techbio'[3]. AI technology has the potential to revolutionise the life sciences and healthcare industry and many other biology-based technology sectors. It is becoming increasingly important in the mission of speeding up the development timeline for much-needed medicines, medical devices and diagnostics, lowering the costs of R&D in the sector, selecting those patients who will benefit most from the treatment, and making diagnostics more accurate.

AI use & data access in techbio

The techbio sector sets itself apart from the creative industries and from 'AI firms' that are at the centre of the ongoing public debate and the considerations of this consultation[4]. Not only because of the use case for AI, but also the types of data that are accessed by techbio companies. Techbio companies will often access data that is subject to strict data protection policies already, such as

[1] DSIT, DHSC, OLS: Bioscience and health technology sector statistics 2021 to 2022. (2023)
[2] ONS: Business enterprise research and development, UK: 2022. (2024)
[3] BIA: TechBio: UK leads innovation frontier. (2024)
[4] IPO, DSIT, DCMS: Copyright and Artificial Intelligence. (2024)

patient data, and needs to be carefully licensed, or data that lies behind paywalls such as publications. Publicly available data, the main subject of this consultation, is also an important component for techbio companies, as such data is fundamental to the training of AI systems and models that techbio companies build upon and develop further for their impactful use case.

The use of AI and technology in the life sciences is central to the Government's ambitions to increase economic growth and to improving the efficiency of the NHS, and access to data is vital for this. It is important for Government to keep in mind the potential impact of the outcomes of the consultation and following policy decisions on UK life sciences SMEs and their ability to continue to harness data and AI for the benefit of the UK's healthcare sector and the public.

Consultation responses

1. Do you agree that option 3 is most likely to meet the objectives set out above?
and
2. Which option do you prefer and why?

Yes, the BIA agrees that option 3 is most likely to meet the objectives on control, access and transparency, and is in line with BIA's response to the 2022 consultation[5]. The current Text and Data Mining (TDM) exception is restricted to "non-commercial use", which is a narrow exception in practice. On the understanding that the majority of AI and machine learning expertise lies in the commercial sector, and given how valuable it is overall to foster R&D for new therapies and diagnostics through the secondary use of data, the BIA welcomes option 3 which allows mining of copyright works, for both non-commercial and commercial use. (The BIA would also support a similar exception for database rights.) In essence, if a user is given lawful access to copyright works or a database, the rights should be the same whether that information is read by a human or a machine. Nevertheless, in order to strike the right balance and to ensure that businesses are incentivised to invest in, build and share good quality data sets, BIA also supports the data holders' rights to expressly reserve their rights on the use of their data.

[5] BIA: BIA submission to the UK IPO's consultation on AI and IP. (2022)

UK BioIndustry Association

3. Do you support the introduction of an exception along the lines outlined above?

and

4. If so, what aspects do you consider to be the most important? If not, what other approach do you propose and how would that achieve the intended balance of objectives?

Yes, the BIA supports the introduction of an exception with rights reservation as outlined by option 3.

Under the current system, it is not always clear whether rights holders do or do not object to their data being mined by, or for the purpose of training, AI models. We support the proposal under option 3 which would shift the current position towards a right to mine, where lawful access exists, with the option for data holders to reserve their rights.

It is important to note that many datasets being used in the life sciences industry include sensitive data such as patient data. In particular, access to pseudonymised patient-level data is becoming increasingly important for the purposes of gaining valuable insights in order to develop effective medicines and diagnostics. In practice, therefore, data privacy rules are likely to heavily constrain many of the benefits that techbio companies and other developers in the life sciences sector would gain from having a wider TDM exception (although some advantages will still remain, e.g. for activities such as web-scraping of chemical literature for the discovery of new molecules).

It is crucial that any new TDM mechanism does not cut across or diminish any rights of privacy. In practice, given the potential complications between the rules on TDM and privacy, the BIA would welcome a joined-up approach between the UK IPO and the Information Commissioner's Office (ICO)[6] to provide guidance to businesses to help navigate between them.

5. What influence, positive or negative, would the introduction of an exception along these lines have on you or your organisation? Please provide quantitative information where possible.

The introduction of an exception as outlined by option 3 would allow techbio companies to access larger, better quality and more diversified datasets through data mining for commercial use. This is key to developing new healthcare solutions with the power of AI, and to reducing the potential problems with bias in AI systems.

Many of the most valuable datasets or works lie in the hands of private companies, including scientific, technical, and medical (STM) publishers of journals for example, some of whom hold works or data that is either difficult or impossible to obtain elsewhere. There are challenges and

[6] ICO: Information Commissioner's Office response to the consultation series on generative AI. (2024)

often high costs associated with negotiating data licences with STM publishers. Negotiating licences with such publishers is often difficult and time-consuming and is becoming increasingly expensive, with multiple layers of different types of access and numerous restrictions around use (and sometimes attaching significant strings to any outputs that a company generates from the use of the publishers' content). Additional fees are usually required for text and data mining access.

BIA are therefore supportive of the proposals to find ways to facilitate and encourage the licensing of data on fair and clear terms, as well as increasing transparency measures to ensure rights holders are aware of where their work is used, for example, to train AI models. Developing model licences for commercial data sharing and associated best practice guidance would be welcomed. These will be very valuable to help standardise and simplify the process and reduce the negotiation costs for less complex arrangements.

There is a risk that any such new exception could be used by STM publishers to make it overly burdensome and costly for AI developers and techbio companies to access scientific works and data. The proposed mechanism that allows rights holders to reserve their rights should be simple and affordable, noting that large parts of the life sciences and biotech industry are startups and SMEs, research institutions and other smaller actors. A new policy mechanism should not be too restrictive or onerous as this will make it too difficult for SMEs to navigate, and lead to disadvantaging smaller, innovative and R&D-driven businesses.

6. What action should a developer take when a reservation has been applied to a copy of a work?

When a reservation has been applied to a copy of work, an AI developer should be able to show that it has undertaken a diligent search to ensure that, to the extent possible, the use of any expression of the reserved work is avoided. As an example, such a diligence requirement could be similar to that required in the context of the orphan works licensing scheme[7] and establishing whether a work is an orphan work.

7. What should be the legal consequences if a reservation is ignored?

If a reservation is ignored, the remedies should be proportionate, and the consequences in line with those currently available for copyright infringement, such as injunctions, damages, delivery up, and legal costs. Springboard injunctions may be appropriate in some circumstances.

[7] IPO: Orphan works diligent search guidance for applicants. (2021).

Aside from AI developers, others also need to be considered and the consequences of the actions of third parties. For example, if a third party deliberately removes a rights reservation in a copyright work, any subsequent reader may not know that such a reservation once existed. It would not necessarily be possible to establish breach of contract (e.g. T&Cs) in those situations as no contract between the relevant parties may exist. Therefore, careful consideration needs to be given to scenarios that may arise, and appropriate legal consequences for the actors involved (inadvertently or otherwise).

8. Do you agree that rights should be reserved in machine-readable formats? Where possible, please indicate what you anticipate the cost of introducing and/or complying with a rights reservation in machine-readable format would be.

Yes, rights should be reserved in machine-readable formats. However, it is vital that it will be implemented as a simple and affordable solution that is easy to navigate and not overly burdensome to startup and SME actors.

9. Is there a need for greater standardisation of rights reservation protocols?

Yes, greater standardisation of rights reservation protocols would be welcomed.

10. How can compliance with standards be encouraged?
and
11. Should the government have a role in ensuring this and, if so, what should that be?

Compliance could be encouraged by, for example, stipulating that certain remedies are only available to rightsholders if they have opted out in the standardised way, and developers will be deemed to have acted appropriately or diligently in checking for rights reservations by checking for the standardised reservations (and possibly checking in the standardised way if that is part of the standard).

Government can play an important role in developing industry guidelines or codes of practice, which are helpful tools to support industry to comply with standards.

14. Should measures be introduced to support good licensing practice?

As noted under Question 5 above, there are existing challenges and costs associated with negotiating data licences with STM publishers. Measures around good licensing practice, such as model licenses for commercial data sharing and associated best practice guidance, would be

welcomed. These will be very valuable to help standardise and simplify the process and reduce the negotiation costs for less complex arrangements, and to facilitate the licensing of data on fair and clear terms.

15. Should the government have a role in encouraging collective licensing and/or data aggregation services? If so, what role should it play?

In the life sciences and biotech sector, there are significant efforts ongoing on enabling techbio companies to access NHS data securely. Joint government-industry schemes such as the NHS Data Access Request Service (DARS) and the Secure Data Environment (SDE) work are looking to enable data aggregation and secure access through controlled environments. These allow access to streamlined data, aggregating data that is currently held by several separate NHS Trusts, all while maintaining data security and privacy. Industry is supportive of such systems, provided they facilitate data access in a practical, secure, and timely way. Similarly in Europe, the European Health Data Space (EHDS) initiative will provide a framework for the secure exchange of health or patient data records across the EU to provide better healthcare, and to provide an improved and more trustworthy basis for the re-use of health data for R&D and innovation.

Such data access and aggregation government initiatives around sensitive health and patient data are welcomed and supported by the biotech industry.

17. Do you agree that AI developers should disclose the sources of their training material?
and
18. If so, what level of granularity is sufficient and necessary for AI firms when providing transparency over the inputs to generative models?

There are several considerations to make on whether or when an AI developer should be asked to disclose the sources of their training material.

- Where an AI developer is using its own data or the data of a collaboration partner, disclosure would not be needed nor necessarily appropriate. For example, when the data used by an AI developer is part of a research collaboration agreement that includes contractual confidentiality restrictions imposed by the collaborator on making public disclosures.

- Considering the transparency objective outlined in the consultation proposal, which aims to inform data holders about the use of their data for training AI models, disclosure of sources would be welcomed *where the data holders would otherwise not be aware of such use*. This

would make it easier for copyright holders to understand how their works have been used, and whether such use was in fact permitted, and monitor potential output infringements.

- Where the data holder has not expressly reserved their rights, disclosure by AI developers would not appear to be necessary, but as discussed above, in situations where an opt-out has not been observed (inadvertently or due to interference by third parties), transparency would enable the rights holder to challenge whether the developer has acted diligently.

When providing transparency over the inputs to generative models, the level of granularity necessary should be at a relatively high level. It should strike the right balance between being useful to copyright holders, while not including a level of technical detail that may jeopardise the AI company's trade secrets, or a level that would become unworkable. There also needs to be consideration as to whether any transparency measures would apply retrospectively, that is, to training material used before the proposed legislative change.

24. What steps can the government take to encourage AI developers to train their models in the UK and in accordance with UK law to ensure that the rights of right holders are respected?

As mentioned above, a simple and affordable mechanism that is easy to navigate and not overly burdensome to startup and SME actors, will play an important part in encouraging AI developers to operate in the UK. In addition, harmonisation, to the degree possible, with the EU legislative position on text and data mining would make it easier and more manageable for AI developers working internationally.

28. Does the existing data mining exception for non-commercial research remain fit for purpose?

No, the existing data mining exception for non-commercial research is not fit for purpose. BIA support option 3, which would extend the current TDM exception to commercial as well as non-commercial use (subject to appropriate safeguards).

See response to questions 1 and 2 above.

29. Should copyright rules relating to AI consider factors such as the purpose of an AI model, or the size of an AI firm?

While smaller actors in the industry should not be disadvantaged through an overly difficult or burdensome mechanism, it would be complex if not unworkable to consider the purpose of an AI model or the size of an AI firm in the copyright rules themselves. While we want to foster the ability

of AI companies to train models to help develop innovative healthcare solutions, this should not be at the expense of data holders in the life sciences industry (such as large pharmaceuticals, biotech SMEs and startups, and research institutes) by making it easier for some AI developers to have greater rights over others for the purposes of scientific or medical research. Relaxing or removing safeguards for some rights holders, such as constraining their ability to reserve their rights depending on the intended use of the data, could create legal uncertainty and have unintended consequences (for example, uncertainty over what would happen to an SME and its AI model if the SME were acquired by a larger company).

30. Are you in favour of maintaining current protection for computer-generated works? If yes, please explain whether and how you currently rely on this provision.

The life sciences sector is typically concerned with computer-generated data and databases. Highly valuable proprietary datasets are created using AI from aggregating, cleansing, processing and analysing existing, usually human-generated data. However, the outputs tend to be unstructured data, more often than copyrightable works. The life sciences sector tends not to rely heavily on the computer-generated works provision, in contrast to some other industries.

From the life sciences sector perspective, the BIA does not see a compelling reason to change the status quo on computer-generated works. The test for who is the author of a computer-generated work (namely "the person by whom the arrangements necessary for the creation of the work are undertaken") could be adopted as a more general principle in relation to AI-generated works and inventions (that is, AI generated IP) and would provide for a more consistent approach.

35. Are you in favour of removing copyright protection for computer-generated works without a human author?

No, the BIA is not in favour of changing the status quo.

46. What are the implications of the use of synthetic data to train AI models and how could this develop over time, and how should the government respond?

Synthetic data is an important and growing initiative being deployed in the life sciences industry. Larger data sets can be developed from real-world data sets, enabling AI providers to train and fine tune their AI models to ensure greater accuracy and precision, and, for example, to generate data to develop treatments for rare diseases where there is a natural shortage of real-world data.

For example, a synthetic control arm can be used in clinical trials of a new drug, where the current treatment is well-established, and the progression of the disease is relatively predictable. In these

cases, there is often a wealth of real-world data that can be used to generate a synthetic dataset for the control arm, making it significantly less expensive for biotech and pharmaceutical companies to conduct clinical trials. Patients can be enrolled and dosed with the new drug, without having the control arm testing patients on the standard treatment. This reduces the number of patients who need to partake in a trial. This route to drug approval has been recognised by the FDA who has produced guidance on the design of such clinical trials, and we can expect other regulators to follow suit.

However, the use of synthetic data can present particular issues with system bias, that is, data taken from a non-representative or biased real-world data will be amplified as such, which needs to be considered.

More detailed, practical guidance on the use of synthetic data in healthcare research, including in relation to dealing with the potential risks of re-identification with the original patient datasets, would be welcomed.

For further information about the contents of this submission, please contact Linda Bedenik, Senior Policy & Public Affairs Manager, via lbedenik@bioindustry.org.