Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR?

*Janos Meszaros, Chih-hsing Ho**

Academia Sinica, Taiwan

ARTICLE INFO

Keywords:
GDPR
AI
Artificial intelligence
Machine learning
Scientific research
Computer science
Commercial research

ABSTRACT

The paper examines how the EU General Data Protection Regulation (GDPR) is applied to the development of AI products and services, drawing attention to the differences between academic and commercial research. The GDPR aims to encourage innovation by providing several exemptions from its strict rules for scientific research. Still, the GDPR defines scientific research in a broad manner, which includes academic and commercial research. However, corporations conducting commercial research might not have in place a similar level of ethical and institutional safeguards as academic researchers. Furthermore, corporate secrecy and opaque algorithms in AI research might pose barriers to oversight. The aim of this paper is to stress the limits of the GDPR research exemption and to find the proper balance between privacy and innovation. The paper argues that commercial AI research should not benefit from the GDPR research exemption unless there is a public interest and has similar safeguards to academic research, such as review by research ethics committees. Since the GDPR provides this broad exemption, it is crucial to clarify the limits and requirements of scientific research, before the application of AI drastically transforms this field.

1. Introduction

Devices using AI and machine learning become part of our lives. What was science fiction before, such as self-driving cars and medical robots in Star Wars, has become reality. However, what was left out from movies is that these devices need to be programmed and trained with large amounts of personal data. The main issue is that this data cannot always be acquired directly from individuals with informed consent. Therefore, the development and research related to these devices are mostly done with the secondary use of personal data. What the movies got right was that these devices can malfunction, which may cost lives. To prevent bias, the data need to be accurate and representative, which can be hindered, if many data subjects choose to opt-out.

Since innovation is crucial for the European Union (EU), the exemptions for scientific research in the EU General Data Protection Regulation (GDPR) permit the reuse of personal data for research purposes. The GDPR defines scientific research in a broad manner, including publicly and privately funded research. This broad exemption aims to provide flexibility to conduct a wide range of scientific research. However, this definition permits private companies to conduct commercial research, and they might not have in place the same level of ethical and institutional safeguards as academic researchers. Furthermore, public interest does not have to be apparent in commercial research. Therefore, commercial AI research should

* Corresponding author at: Chih-hsing Ho, Institute of European and American Studies, Academia Sinica, 128, Sec 2, Academia Road, Nankang, Taipei, 115, Taiwan.
E-mail address: chihho@sinica.edu.tw (C.-h. Ho).

https://doi.org/10.1016/j.clsr.2021.105532
0267-3649

not fit into the GDPR research exemption without public interest and similar safeguards as academic research.

In order to shed more light on these issues, Part II examines the Google DeepMind and Cambridge Analytica cases to highlight concerns about the sharing of personal data for AI research. Part III elucidates the regulation of scientific research in the GDPR. Part IV tackles the question of whether the prohibition on solely automated decision-making in the GDPR poses a significant hurdle for AI research. The next part highlights the differences between academic and commercial research, and Part VI focuses on the development of AI products and services. The final parts (VII and VIII) elucidate how the development of AI products and services fits into the definition of scientific research in the GDPR. The paper concludes with feasible solutions to find a balance between interests in privacy and innovation.

2. Sensitive data for AI research: the Google DeepMind and Cambridge Analytica cases

In 2016, Google DeepMind[1] and the Royal Free London NHS Foundation Trust ("Royal Free")[2] signed a data-sharing agreement. The Royal Free provided the healthcare records of 1.6 million patients to develop 'Streams,' an AI diagnosis application for acute kidney injury. However, there have been several issues with this data-sharing agreement, such as the lack of approval from relevant authorities (e.g., Health Research Authority [HRA] and Information Commioner's Office [ICO]), and inadequate processes to inform the data subjects.[3] The large volume of records containing sensitive health data was not de-identified, as the Royal Free believed that the data was being processed for the purpose of direct patient care; thus, the parties did not have explicit patient consent either.

The ICO[4] launched an investigation in 2017 and found that the Royal Free failed to comply with the Data Protection Act when it provided patient details to Google DeepMind.[5] Furthermore, the ICO ruled that Royal Free did not have a valid legal basis for satisfying the common law duty of confidentiality and therefore, the processing of the data breached that duty.[6] The ICO clarified that the purpose of data processing was not direct care, and inferred that it was rather research, development or clinical improvement.[7] The National Data Guardian (NDG) was also of the opinion[8] that the 'purpose for the transfer of 1.6 million identifiable patient records to Google DeepMind was for the testing of the Streams application, and not for the provision of direct care to patients.' After the investigation, DeepMind and Royal Free were able to continue their partnership, with additional safeguards: they had to establish a proper legal basis for the data processing, complete a privacy impact assessment and adequately inform the public about the project. After finding a proper legal base and finalising the 'Stream' application, it became clear that the purpose of the processing was research and development with public interest to improve healthcare experiences. However, the degree of public interest, especially in the further use of the data and intellectual property is debatable.

A scandal in 2018 revealed that 87 million Facebook users' personal data was harvested without consent by Cambridge Analytica, a company that aimed to target and manipulate users during political campaigns.[9] However, the Cambridge Analytica's targeting technology originated from the work of researchers at the Psychometrics Centre at Cambridge University. There was a 'close working relationship between Facebook and individual members of the research community', while the Psychometric Centre used Facebook data with AI and machine learning for developing personality profiles.[10] The UK ICO investigation highlighted that academic studies and the commercial enterprises set up by academics could become entangled.[11] The Facebook - Cambridge Analytica case has demonstrated that research results developed with oversight can turn into 'tools' for unethical and unlawful purposes, such as manipulating people.

The two cases highlight the issues around sharing personal data for research purposes, and the narrow line between scientific research and commercial activity. Even with the involvement of public organisations in both cases, research ethics did not prevail. In the case of companies and private research institutions, this issue might be even more concerning. Therefore, clear and strict regulation on scientific research would be crucial.

[1] DeepMind Technologies is a British artificial intelligence company founded in 2010, currently owned by Google through Alphabet Inc.

[2] Royal Free is one of the largest healthcare providers in Britain's publicly funded National Health Service (NHS).

[3] Julia Powles and Hal Hodson, 'Google DeepMind and healthcare in an age of algorithms' (2017) vol. 7 (4) Health and technology, 351-367

[4] The ICO is the UK's independent body set up to uphold information and privacy rights.

[5] Royal Free - Google DeepMind trial failed to comply with data protection law https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/ accessed 15 Sept 2018

[6] ICO, Decision and letter to Royal Free NHS Foundation Trust, RFA0627721 – provision of patient data to DeepMind, 3 July 2017 https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf accessed 13 Aug 2018

[7] Ibid 1.'*First and foremost, my office has made our support for the appropriate use of personal data for the purpose of research, development and clinical improvements clear.*'

[8] Dame Fiona Caldicott, the National Data Guardian wrote this opinion in a letter in 2017 to Professor Stephen Powis, the medical director of the Royal Free Hospital in London, which provided the patients' records to Google DeepMind. The letter was leaked to Sky News in 2017 February.S Alexander J Martin, 'News - Google received 1.6 million NHS patients' data on an 'inappropriate legal basis' (*Sky News,* 15 May 2017) https://news.sky.com/story/google-received-1-6-million-nhs-patients-data-on-an-inappropriate-legal-basis-10879142 accessed on 21 May 2019

[9] J. Isaak and M. J. Hanna, 'User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection' (2018) Computer Science, vol. 51, no. 8, 56-59

[10] ICO Report, *Investigation into the use of data analytics in political campaigns* (2018) 38

[11] Information Commissioner's Office (ICO), *Investigation into the use of data analytics in political campaigns*' (Report to the Parliament of 6 November 2018) 55-58

3. AI research in the GDPR

Leaders of AI research have taken different approaches to developing their products. In the United States, private companies, such as Google and Facebook, dominate the fields of AI research. In China, the government has a strong influence on research with fewer regulatory barriers than in the EU and US.[12] In the EU, research groups operating in different countries with diverse research and legal environments may find their efforts at collaboration hindered.[13] EU level research plans and funds, such as the Horizon 2020, are intended to accelerate both cooperation and competition among researchers. However, the different legal expectations and exemptions for scientific research in the EU may pose hurdles for AI research.[14] The EU has recognised the importance of AI research and the European Parliament has voted on a resolution to regulate the development of AI and robotics across the EU. The proposed rules include establishing ethical standards for the development of AI and the regulation of liability.[15] From the data protection law's point of view, scientific research is a distinguished type of data processing in the GDPR, and the Regulation provides three categories of exemptions for research:

1) Exemptions from data processing principles and lawful grounds for processing;
2) Exemptions from the data subjects' rights;
3) The Member States can implement further research exemptions.

3.1. *Exemptions from data processing principles and lawful grounds for processing*

One of the data processing principles in the GDPR is the purpose limitation, which means personal data cannot be further processed in a manner that is incompatible with the original purpose. For instance, when personal data is collected for healthcare purposes, it cannot be used for marketing. However, the GDPR recognises it is often not possible to fully identify the purpose of processing for scientific research at the time of data collection.[16] Therefore, the GDPR provides an exemption from the purpose limitation principle: further processing for scientific research purposes deemed to be compatible with the initial purposes, if safeguards are satisfied.[17] The European Data Protection Supervisor (EDPS) clarified the meaning of this presumed compatibility: 'The presumption is not a general authorisation to further process data in all cases for historical, statistical or scientific purposes. Each case must be considered on its own merits and circumstances. But in principle personal data collected in the commercial or healthcare context, for example, may be further used for scientific research purposes, by the original or a new controller, if appropriate safeguards are in place.'[18] Similarly, the UK Medical Research Council interpreted this exemption as the 'GDPR says any personal data can be used for research, regardless of the initial reason for collection, subject to safeguards, transparency and fairness.'[19]

The European Data Protection Board (EDPB) is of the opinion that the secondary compatible use for scientific research might not need a separate legal basis by stating '…the controller could be able, under certain conditions, to further process the data without the need for a new legal basis.' However, the EDPB highlighted that this exemption would require specific attention and guidance in the future.[20] The EDPS approaches this issue more carefully by promoting a compatibility test for the reuse of data, particularly in the context that the data was originally collected for very different purposes or outside the area of scientific research.[21] These documents from the EDPB and EDPS clarify that they have no intention to interpret the GDPR against the lawmakers' will; thus, secondary use should be lawful. However, the authorities also cannot authorise the secondary use of data without limitation since research cannot be a carte blanche to take irresponsible risks. Therefore, the authors argue that the reuse of data should only be allowed after a compatibility test, especially in the case of data originally collected for public interest purposes. The GDPR introduces a compatibility test under Article 6 (4), to consider several circumstances: 1) any link between the new and original purpose; 2) the context in which the personal data have been collected; 3) the nature (sensitivity) of the personal data; 4) possible consequences of the intended further processing; 5) the safeguards to protect the data. The data controllers are responsible for conducting this test. However, the GDPR does not specify how this test needs to be performed, documented, and overviewed. The authors argue that data controllers should conduct this test based on EU level guidelines, and the results of it should be reviewed by relevant authorities. Since the GDPR mandates the data protection authorities (DPAs) to force out its requirements, the easiest way to review the secondary use of data might be the overview by national DPAs, with the help of authorities or related ethics committees responsible for scientific research. In the corporate environment, for privately funded research that solely serves private interest, the compatibility test needs to conform to strict standards, and approved by authorities.

[12] Centre for Data Innovation, *Who Is Winning the AI Race: China, the EU or the United States?* (2019)
[13] Dove, E. S., 'The EU General Data Protection Regulation: Implications for International ScientificResearch in the Digital Era' (2018) 46(4) The Journal of Law, Medicine & Ethics, 1013-1030
[14] Humerick, Matthew, 'Taking AI Personally: How the E.U. Must Learn to Balance the Interests of Personal Data Privacy & Artificial Intelligence' (2018) 34 Santa Clara High Tech. L.J., 393
[15] Communication from the Commission to the European Parliament, the European Council, The Council, The European Economic and Social Committee and the Committee of the Regions, *Artificial Intelligence for Europe* (Brussels, 2018)
[16] See Recital 33, GDPR
[17] See Article 5 (1) b, GDPR
[18] European Data Protection Supervisor, *Preliminary Opinion on data protection and scientific research* (2020) 22
[19] UK Medical Research Council, GDPR and Data Protection Act 2018: Key facts for research, p. 1.https://mrc.ukri.org/documents/pdf/gdpr-key-facts-for-research/
[20] European Data Protection Board, *Opinion 3/2019 concerning the Questions and Answers onthe interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (art.70.1.b)* (23 January 2019)
[21] European Data Protection Supervisor (2020) 23

However, currently, there is no common understanding in the EU, how to regulate and approve these activities. Compared to statutory regulation, codes have the advantage of faster and more specific shaping of technology.[22] However, as Raab highlights, recent codes addressing AI and machine learning still operate with 'headline values', such as transparency, respect for human dignity and autonomy, lacking the specific guidance. Furthermore, the development and application of AI devices involve several actors in the economy, society and politics, and it is not straightforward who is responsible for the implementation and enforcement of these values. As Amram and Ho point out, data protection compliance walks together with the ethical one.[23] Therefore, the close collaboration between authorities responsible for data protection and overviewing scientific research could solve this issue.

3.2. *Exemptions from the data subjects' rights*

Since the GDPR is a regulation, data controllers can directly rely on several exemptions without the need for these provisions to be implemented into national legislation. For instance, in the case of scientific research, the request for erasure (right to be forgotten) can be rejected,[24] if it is likely to render impossible or seriously impair the achievement of research purposes. For scientific research with public interest, the right to object can be also exempted,[25] as in these cases, the research may not yield reliable results because of the objection or erasure request; or in the worst case, the research cannot be started or completed because of the prohibitive costs and administrative burden.[26] As Ducato highlights, the erasure would also risk undermining the scientific validity of the research by preventing the verification of the results and hinder the peer-review process.[27] Determining the scope of these exceptions requires balancing a number of separate considerations, which poses a challenge for researchers without detailed regulation and official guidelines.[28]

3.3. *Research exemptions implemented by the EU Member States*

The final category of research exemptions can be implemented by the EU Member States into national law pursuant to Articles 89 (2) and (3) of the GDPR. Data subjects have several rights that may be exercised to exert control over the processing of their personal data, such as the right to access and rectification.[29] However, the GDPR allows the Member States to decide if many of these rights can be applied in the case of scientific research,[30] if the application of them would impede or render research impossible.[31] Thus, the application of these derogations are limited, and they can only be applied if appropriate technical and organisational safeguards are in place pursuant to Article 89 (1). Some Member States have already introduced derogations. For instance, the German New Federal Data Protection Act (FDPA) limits the data subjects' rights in the context of scientific research: individuals cannot assert their rights of access, correction, restriction and objection if it would make the scientific research impossible or cause serious impairment in it.[32] However, these derogations are against another main goal of the GDPR, which is the standardisation of data protection rules in the EU. Therefore, the implementation of these derogations has caused significant concern within the research community.[33] The main reason for allowing derogations for Member States is the lack of conferred competency of the EU in this field.[34] Extraordinary situations, such as the COVID-19 outbreak, highlight that restrictions in the rights of data subjects vary greatly among the EU Member States.[35]

---

[22] Charles D. Raab, Information privacy, impact assessment, and the place of ethics, Computer Law & Security Review, Volume 37, 2020, 105404,ISSN 0267-3649, https://doi.org/10.1016/j.clsr.2020.105404.

[23] Denise Amram, Building up the "Accountable Ulysses" model. The impact of GDPR and national implementations, ethics, and health-data research: Comparative remarks, Computer Law & Security Review, Volume 37, 2020, p. 7.Chih-hsing Ho, 2018, Challenges of the EU General Data Protection Regulation for Biobanking and Scientific Research, Journal of Law, Information and Science, Volume 25, Issue 1, EAP. 1-20

[24] See Article 17 (3) d), GDPR

[25] See Article 21 (6), GDPR

[26] M. E. Kho et al., 'Written Informed Consent and Selection Bias in Observational Studies Using Medical Records: Systematic Review' (2009) BMJ, 338C. Junghans, M. Jones, 'Consent bias in research: how to avoid it' (2007) 93(9) Heart 1024Rothstein MA, Shoben AB., 'Does consent bias research?' (2013) 13(4) Am J Bioeth, 27

[27] Ducato, Rossana, Data Protection, 'Scientific Research, and the Role of Information' (2020) Computer Law and Security Review, 6

[28] Politou, E., Michota, A., Alepis, E., Pocs, M., & Patsakis, C., 'Backups and the right to be forgotten inthe GDPR: An uneasy relationship' (2018) 34(6) Computer Law & Security Review, 1247-1257Pormeister, K., 'Genetic data and the research exemption: is the GDPR going too far?' (2017) 7(2)International Data Privacy Law, 137-146

[29] Van Ooijen, I., & Vrabec, H. 'Does the GDPR Enhance Consumers' Control over Personal Data? An Analysis From a Behavioural Perspective' (2018) 42 Journal of consumer policy, 91–107

[30] Janos Meszaros, Chih-hsing Ho, 'Big data and scientific research: the secondary use of personal data under the research exemption in the GDPR' (2019) Acta Juridica Hungarica, 403-419

[31] See Article 89 (2), GDPR

[32] FDPA Section 27 (2) The rights of data subjects provided in Articles 15, 16, 18 and 21 of Regulation (EU) 2016/679 shall be limited to the extent that these rights are likely to render impossi-ble or seriously impair the achievement of the research or statistical purposes, and such limits are necessary for the fulfilment of the research or statistical purposes.

[33] See e.g., Mourby, M., Mackey, E., Elliot, M., Gowans and others, 'Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK' (2018) 34(2) Computer Law & Security Review, 222-233Timmers, M., Van Veen, E.-B., Maas, A. I. R., & Kompanje, E. J. O., 'Will the Eu Data Protection Regulation 2016/679 Inhibit Critical Care Research?' (2018) Volume 27, Issue 1 Medical Law Review, 59–78

[34] Gauthier Chassang, 'The Impact of the EU General Data Protection Regulation on Scientific Research' (2017) 11 Ecancermedicalscience, 709

[35] For the comparison of different EU and Member State reactions to COVID-19, see:
European Parliament: States of emergency in response to the coronavirus crisis: Situation in certain Member States, 2020, https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/649408/EPRS_BRI(2020)649408_EN.pdf
Deloitte: Privacy and Data Protection in the age of COVID-

Another issue, which requires further clarification and standardisation, are the expected implementation of organisational and technical safeguards, pursuant to Article 89. The GDPR promotes pseudonymisation and anonymisation, as appropriate safeguards. Anonymous data cannot identify individual data subjects; therefore, it is not considered personal data anymore.[36] Pseudonymization is the separation of data from the direct identifiers (e.g., name, address), so that re-identification is not possible without additional information (the 'key') that is held separately.[37] However, as Amram highlights, technically, it does not exist a unique criterion of anonymization, neither pseudonymisation. Therefore, the national implementations of these safeguards may also be differing.

4. Profiling and automated decision-making

The prohibition on solely automated decision-making in the GDPR may pose a significant hurdle for the application of AI in various fields. However, scientific research might be less affected, since mostly the main goal of research activities is producing new knowledge, rather than making decisions for individuals. It is crucial to differentiate between profiling and solely automated decision-making, since profiling without solely automated decision-making is not prohibited by the GDPR.[38] The GDPR defines profiling as the automated processing of personal data to analyse or make predictions about individuals. For instance, predicting performance at work or personal preferences.[39] Profiling is composed of two main elements: 1) automated processing, which does not have to be solely automated; and 2) the purpose is to evaluate personal aspects about an individual. Profiling can be one of the sources for automated decision-making. In the case of speeding tickets, when the police automatically fine drivers based on data collected from a traffic camera system, it is automated decision-making based on observed data. However, when citizens' driving habits are evaluated to calculate their fines, such as previous offences, then the automated-decision is based on profiling.[40]

The Article 29 Working Party clarified that the GDPR prohibits solely automated decision-making, not just provides an opt-out. The solely automated decision-making in the GDPR consists of three main parts: (1) a decision must be or has been made; (2) that decision is based solely on automated processing; (3) the decision has either legal effects or similarly significant consequences.[41]

Since the main goal of scientific research is to produce knowledge, the first element (decision on individuals) is usually not fulfilled; thus, the prohibition does not exist. However, it is possible that research produces decisions based on individual data. For instance, if an AI application analyses X-ray pictures, decides which patient might have a high chance of cancer and needs further medical examination. The second element (solely automated processing) is fulfilled if the decision made from scientific research is solely automated. In the previous example, if the AI application makes a decision alone based on the X-ray images, then it is solely automated. However, if the application provides information for a radiologist, and she makes the final decision, then it is not a solely automated decision, since the AI application only assists the doctor, and there is meaningful human involvement. The third element (legal effects or similarly significant consequences) might be something that affects a person's legal status or their rights significantly, such as influence on health or finance.[42]

Data subjects have several rights, which might not be applied in the case of scientific research, based on Union or Member State law (e.g., right to access, rectification, erasure, restriction of such processing, and the right to object). Since automated decision-making is not mentioned in this part, scientific research cannot avoid this prohibition. The reason for it might be the goal of scientific research, which is to produce new knowledge without direct consequences for individuals.[43] In the case researchers make decisions, there are ways for them to avoid the GDPR's restrictions on automated decision-making. For instance, with meaningful human involvement in the decision process. Therefore, using profiling without solely automated decision-making is not prohibited, which is crucial for AI research. However, authorities and related research ethics committees need to pay careful attention to how decisions are made within the scope of the projects.[44]

5. Artificial intelligence and scientific research

5.1. *Scientific research in the European Union*

It is a common practice in scientific research to process personal data for a different purpose than the original one to pursue new findings with the same dataset to save time and re-

19, 2020 https://www2.deloitte.com/content/dam/Deloitte/be/Documents/risk/be-risk_privacy-and-data-protection-in-the-age-of-covid-19.pdf

[36] See Recital 26, GDPR

[37] See Article 4 (5) GDPR

[38] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profilingfor the purposes of Regulation 2016/679* (2018) 19

[39] GDPR Article 4 (4) 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;

[40] Article 29 Working Party on Profiling (2018) Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2018) 19.

[41] GDPR Article 22 (1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

[42] Article 29 WP, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2018) 21

[43] Ducato, Rossana, Data Protection, 'Scientific Research, and the Role of Information' (2020) Computer Law and Security Review, 14

[44] European Parliamentary Research Service, *How the General Data Protection Regulation changes the rules for scientific research* (2019) 34

sources.[45] The GDPR recognises it is often not possible to fully identify the purpose of processing for scientific research at the time of data collection.[46] This statement is crucial, since obtaining consent became more challenging under the Regulation, which must be unambiguous and specific.[47] Scientific research has a distinguished position in the GDPR. For instance, personal data can be further processed for research,[48] and several rights of the data subjects can be bypassed (e.g., right to erasure).[49] Hence, it would be crucial to have a comprehensive and legally binding definition of scientific research in the EU, clarifying the types of activities that could qualify as scientific research. In the case of processing sensitive data for research purposes, the GDPR requires the processing to be based on Union or Member State law, which protects the data subjects and proportional to the achievable purpose.[50] As the European Data Protection Supervisor (EDPS) highlighted, many Member States have not enacted such laws yet.[51] However, in the age of the COVID-19 pandemic, it is not clear if the enacted laws regulating data protection during an emergency might be a proper tool for this purpose in the long run.[52] After the pandemic, it would be crucial to re-evaluate this issue.

There is no universally agreed-upon definition of scientific research. The GDPR defines it in a broad way, stating that "processing of personal data for scientific research purposes should be interpreted in a broad manner, including, for example, technological development and demonstration, fundamental research, applied research and privately funded research."[53] However, this definition is in the recital part of the GDPR; thus, it is not legally binding. Furthermore, there is no EU law that comprehensively regulates the definition and requirements of scientific research; thus, the authors contacted the European Data Protection Board and all the national Data Protection Authorities (DPAs) in the EU to find an answer as to how 'scientific research' is regulated in their countries, with regard to the GDPR research exemption. From the answers, it became clear that most of the DPAs do not have a special interpretation of the definition of scientific research in the GDPR; they follow the regulations and decisions of the authorities responsible for scientific research in their countries (e.g., Ministries of Science and Education). In our research contacting with DPAs in the EU Member States, eleven DPAs mentioned that there are specific regulations for scientific research in their countries (Italy, Germany, France, Belgium, Ireland, Czech Republic, Sweden, Bulgaria, Slovenia, Poland, Finland), and six respondents answered that there is no specific regulation in their countries (Lithuania, Portugal, Romania, Croatia, Luxembourg, Malta). The DPA in Malta further pointed out scientific research is generally permitted on the basis of public interest. However, the tension between data protection and scientific research was emphasised by the Irish DPA: 'data protection law is heavily context-driven' and 'it is not within the remit' of the data protection authorities to advise or comment on legislative provision outside data protection.'

5.2. *The difference between scientific research and statistical purposes*

AI development often consists of the creation of statistical models; therefore, it is crucial to differentiate between scientific research and statistical purposes. They are two different types of processing in the GDPR.[54] However, statistical and scientific research purposes might overlap, since researchers can use statistics as one of the means to advance knowledge.[55] To differentiate between them, it is crucial to make a distinction between the individual and collective use of personal data. Individual use has a personalised purpose for a particular person (e.g., measuring the individual's daily activity). Collective use relies on personal data with the concentration into a statistical result, which is anonymous data. When the GDPR operates with statistical purposes, it requires the application of the collective use of data. The GDPR defines statistical purposes as the collection and the processing of personal data necessary for statistical surveys or the production of statistical results. The most important characteristic of statistical processing is that the result of processing should be aggregate data, which is not used to support measures or decisions regarding any particular natural person.[56] Thus, the statistical confidentiality[57] and aggregation of data together protect the data subjects from being re-identified.[58] For instance, a company is able to use statistical methods to measure the overall satisfaction of its clients under the GDPR. However, if the company identifies the clients individually through its statistical

[45] Auffray C, Balling R, Barroso I, 'Erratum to: Making sense of big data in health research: towards an EU action plan' (2016) 8(1) Genome Medicine, 118

[46] See Recital 33 and 65, GDPR

[47] See Article 4 (11), GDPR

[48] See Article 9 (2) (j), GDPR

[49] See Article 17 (3) (d), GDPR

[50] See Article 9 (2) (j), GDPR: processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

[51] European Data Protection Supervisor, 'A Preliminary Opinion on Data Protection and Scientific Research' (2020), 23.

[52] Gianclaudio Malgieri, Data protection and research: A vital challenge in the era of COVID-19 pandemic, Computer Law & Security Review, Volume 37, 2020, 3,

[53] See Recital 159, GDPR

[54] See Recital 62, GDPR

[55] Council of Europe, *Explanatory Memorandum. Recommendation No.R (97) 18 of the Committee of Ministers to Member States concerning the protection of personal data collected and processed for statistical purposes* (1997) 6

[56] See Recital 62, GDPR

[57] Regulation 223/2009 Article 2(e) defines statistical confidentiality as the "protection of confidential data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources and implying the prohibition of use for non-statistical purposes of the data obtained and of their unlawful disclosure"

[58] Ducato, Rossana, Data Protection, 'Scientific Research, and the Role of Information' (2020) Computer Law and Security Review, 14

Table 1 – The comparison of academic and commercial research.[59]

	Academic research	Commercial research
Focus	Basic and applied research	Applied research
Basic rationale	Advance knowledge	Increase efficiency
Aim	New ideas	Profit
Characteristics	Idea-centred	Practical, product-centred
Schedule	Open-ended, longer periods	Tight, predetermined
Recognition	Scientific honours	Payment
Framework	Open	Close, confidential
Evaluation	Peer-review	By the leaders
Dissemination	Through academic publishing (e.g., journals, books)	Internal reports, professional conferences
Oversight	Rigorous institutional safeguards (e.g., ethics committee)	Less strict overview

methods, that does not fit in the GDPR's statistical purposes exception.

5.3. *Academic and commercial research*

Research can take the form of basic research or applied research. Basic or fundamental research is an experimental or theoretical work undertaken primarily to acquire new knowledge, without pursuing commercial value. However, results may later lead to the materialisation of a practical product or service. Applied research is 'the planned research or critical investigation aimed at the acquisition of new knowledge and skills for developing new products, processes or services, or for bringing about a significant improvement in existing products, processes or services.'[60] This type of research is designed to solve practical problems of the modern world (e.g., how to make computers faster, batteries lasting longer).

Both basic and applied research can fall within academic and commercial research. The ultimate goal of commercial research is gaining profit; thus, it is more practical, and it prefers applied research (Table 1). Academic research has more focus on basic research with the aim of advancing general knowledge, with public interest. However, there is a growing pressure on universities and public research institutions to adopt a more entrepreneurial approach, resulting in commercialisation in the traditional basic knowledge generation.[61] For instance, public-private partnerships (PPPs)[62] are one of the forms of this commercialisation, gaining increased popularity in the EU. PPPs have been touted as a mutually beneficial financing mechanism. However, corporations usually have legal and fiduciary duties to maximise profits and shareholder returns. These duties and incentives can come at the expense of public interest, particularly when the leaders of corporations are incentivised to link their payment and bonuses to these returns. In this environment, it is more challenging to differentiate between academic and commercial research. Yet, the main differences are the openness and overview of research, as well as whether the public benefit exists. Corporate secrecy is a barrier for independent researchers and authorities to validate and conduct oversight, which is essential for accountability. In their privacy policies, companies allow themselves with vague terms to further use personal data for research purposes. However, this processing is not transparent and does not have to serve public interest purposes. The Cambridge Analytica scandal highlighted how far research could go from reasonable ethical standards. Furthermore, companies close out independent researchers, reducing oversight. For instance, Facebook restricted access to its application programming interface data in 2018; thus, independent researchers could not analyse the connection among profiles, hate speech, and misinformation.

Wagner draws our attention to the issue of accountability in connection with scientific research.[63] He argues all research should be consistent with the GDPR accountability principle (e.g., with ethical review)[64] and the requirement of privacy by design, ensuring a sound legal basis for developing accountable GDPR compliant scientific research. As Table 1 highlights, accountability might be a concern in the case of commercial research. Therefore, in general, academic research aligns better with the accountability principle of the GDPR. Since the GDPR does not differentiate between academic and commercial research, seemingly the same rules apply to them. However, as Table 1 shows, the safeguards and overview of the two types of research are not the same. Therefore, in general, they pose a different level of risk for the data subjects. The authors argue that the same benefits (e.g., the GDPR research exemption) should only apply to academic and commercial research,

[59] Kalantaridis, C., Küttim, M., 'University ownership and information about the entrepreneurial opportunity in commercialisation: a systematic review and realist synthesis of the literature' (2020) J Technol TransfConceição Vedovello, 'Firms' R&D Activity and Intensity and the University–Enterprise Partnerships, Technological Forecasting and Social Change' (1998) Volume 58, Issue 3, 215-226

[60] European Commission, *Community Framework for State Aid for Research and Development and Innovation* (2006/C 323/01), 2.2 f)

[61] Ahoba-Sam, R., Charles, D., 'Building of Academics' Networks—An analysis based on Causation and Effectuation theory' (2019) Rev Reg Res 39, 143–161

[62] Definition of PPP: An arrangement where the private sector supplies assets and services that traditionally have been provided by the government.International Monetary Fund, *Public-Private Partnerships* (2014) 4https://www.imf.org/external/np/fad/2004/pifp/eng/031204.pdf

[63] Ben Wagner, 'Accountability by design in technology research' (2020) Computer Law & Security Review, Volume 37, 10.

[64] See Article 5 (2), GDPR

if both also have the same requirements: public interest and strong safeguards. Therefore, the Member States need to have suitable regulation to validate and review scientific research, especially commercial research.

To achieve this, it would be crucial to differentiate between academic and commercial research. A great example of this is the new EU Copyright Directive, which recognises that 'despite different legal forms and structures, research organisations in the Member States generally have in common that they act either on a not-for-profit basis or in the context of a public-interest mission recognised by the State. This public-interest mission could be reflected through public funding or through provisions in national laws or public contracts.' The Directive recognises that academic and commercial environments might be tangled due to the commercialisation of research. Therefore, the Directive highlights that organisations upon which commercial undertakings have a decisive influence allowing to exercise control through their shareholders or members, which could result in preferential access to the results of the research due to the structural limitations, should not be considered research organisations for the purposes of the Directive.

The Copyright Directive defines[65] 'research organisation' as a '*university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research:*

(a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or
(b) pursuant to a public interest mission recognised by a Member State; in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation;'

Defining research organisations and differentiating among them would be a feasible way to fairly apply the GDPR research exemption. The role of scientific research is understood to provide knowledge that can in turn 'improve the quality of life for a number of people and improve the efficiency of social services'.[66] The European Data Protection Supervisor highlighted that the specific rules on scientific research in the GDPR reflects a clear intention for at least a minimum level of public interest.[67] Furthermore, prominent research organisations, such as the BBMRI-ERIC,[68] suggested also that the GDPR research exemption should be restricted to public interest research.[69] This public mission and strong safeguards are crucial factors for building trust in the development of future AI technology.

6. Is developing AI products and services scientific research?

6.1. *Computer science*

It might be challenging to identify research activities and differentiate among them in computer science. An example of basic research in computer science is searching for alternative methods of computation, such as quantum computation and quantum information theory. On the other hand, developing a new programming language or operating system might constitute applied research. The development of new applications and substantial improvements in existing software represents experimental development.[70] However, it is challenging to identify the research component in software development. Therefore, an upgrade, addition or change to an existing program or system may be classified as research only if it results 'in an increase in the stock of knowledge.'[71]

For a software development project to be classified as research, its completion must be dependent on a scientific advance, and the aim of the project must be the systematic resolution of a scientific and technological uncertainty. Furthermore, the OECD Frascati Manual clarifies that if the research and development are associated with software as an end product or embedded in that product, it could also be classified as research. For instance, research and development in software can be the development of new operating systems or programming languages, efforts to resolve conflicts within hardware or software based on the process, re-engineering a system or a network, or the creation of new, more efficient algorithms based on new techniques.[72] On the other hand, routine software-related activities do not constitute research. For instance: adding user functionality to existing application programs, the customisation of the software for a particular use, unless 'during this process knowledge is added that significantly improves the base program'.[73] The definitions of the OECD Frascati Manual clarified that the development of software might constitute scientific research, as a component of computer science. However, identifying these activities and applying the GDPR research exemption on them requires careful consideration.

6.2. *AI research as scientific research*

There is no precise, universally accepted definition of AI. As recently defined by the EU Commission, 'AI refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.'[74] The High-Level Expert Group on

[65] Article 2 (1), Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

[66] See Recital 157, GDPR

[67] European Data Protection Supervisor, *Preliminary Opinion on data protection and scientific research* (2020) 11.

[68] Biobanking and BioMolecular Resources Research Infrastructure-Europe Research Infrastructure Consortium.

[69] Shabani, M., & Borry, P., 'Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation' (2018) European Journal of Human Genetics, 26(2), 149

[70] Frascati Manual (2015) 53.

[71] Frascati Manual (2015) 66.

[72] Frascati Manual (2015) 66.

[73] Frascati Manual (2015) 66.

[74] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic

Artificial Intelligence also highlighted that, AI is a scientific discipline, as it includes several approaches and techniques, such as machine learning, machine reasoning, and robotics.[75]

Artificial intelligence already outperforms human intelligence in many cases. For instance, AI has emerged as the champion in a wide range of games.[76] However, people's standards and expectations for AI and machines are rising; thus, a world champion chess program and Apple's Siri might not seem as impressive today as they once did. As John McCarthy lamented, "As soon as it works, no one calls it AI anymore."[77] AI is already applied in a variety of ways, from factory robots to advanced toys, and from speech recognition systems to medical research. John McCarthy coined the term 'Artificial Intelligence' and distinguished between basic and applied research in AI,[78] calling for more basic research to reach human-level intelligence. Herbert A. Simon asserted that AI is an experimental science.[79] As Nick Bostrom points out, the line between artificial intelligence and software, in general, is not sharp.[80] Many applications might be viewed as a generic software rather than AI applications, which refers back to McCarthy's dictum when something works, it is no longer called AI.

We may also define AI as a branch of computer science concerned with the properties of intelligence by synthesising intelligence.[81] Allen Newel and Herbert A. Simon described computer science as "the study of the phenomena surrounding computers. The machine - not just the hardware, but the programmed, living machine - is the organism we study."[82] The early development of AI depended on the rapid progress of hardware, while recently there has been a greater focus on software, and in the AI community machine learning is seen as the most promising way to improve AI research.[83] Machine learning, which is a subset of artificial intelligence, can be supervised, unsupervised or semi-supervised. If supervised, all data are labelled, and the algorithms learn to predict the output from the input data; thus, it is easier for programmers to understand the AI research and development in this case. On the other hand, when machine learning is unsupervised, all data are unlabelled, and the algorithms need to learn the structure from the input data on their own. In most cases, machine learning is semi-supervised, which means that part of the data is labelled, but most of it is unlabelled.[84]

The combination of AI and big data will have a significant impact on data subjects, as almost every aspect of the citizens' lives become subject to predictive applications, such as travel time, work efficacy, health status, and political opinion. AI research combines certain specific characteristics, such as complexity, autonomous behaviour, data-driven, and openness. AI's data-driven characteristics and openness make it especially important from the point of view of data protection law. As the previous points and the High-Level Expert Group on Artificial Intelligence[85] highlighted, computer science is a branch of science, and AI research is a part of computer science; thus, AI research can also be qualified as scientific research. With this qualification, AI research might benefit from the GDPR research exemptions.

7. Discussion

The GDPR regulates scientific research with the expectation of already good scientific practice, such as institutional safeguards, applying the same rules on academic and commercial research. However, as Part V highlighted, commercial research does not have the same level of safeguards and overview, as academic research. Furthermore, the opaque algorithms in AI research and corporate secrecy constitute barriers for oversight. To address these issues, the authors suggest the following measures to protect data subjects without hindering innovation:

(1) The harmonised application of GDPR research exemption on AI research in the EU;
(2) Commercial AI research should not benefit from the GDPR research exemption without public interest and similar safeguards as academic research;
(3) Oversight and enforcement by the EU and Member State authorities (e.g., DPAs and related authorities responsible for scientific research) from the start of AI research until the application of final products and services.

Applying the GDPR research exemption on AI research would be a feasible solution to foster innovation in the EU. However, this technology poses significant risks for data subjects; thus, organisational and technological safeguards are

and Social Committee and the Committee of the Regions,Artificial Intelligence for Europe, COM/2018/237 final

[75] High-Level Expert Group on Artificial Intelligence, A definition of AI: Main capabilities and scientific disciplines, 2019, p. 8.

[76] G. Synnaeve and P. Bessière, 'Multiscale Bayesian Modeling for RTS Games: An Application to StarCraft AI in IEEE Transactions on Computational Intelligence and AI in Games' (2016) vol. 8, no. 4, 338-350David Silver, Thomas Hubert, Julian Schrittwieser at al., 'A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play' (2018) Science, 1140-1144

[77] Vardi, Moshe Y., 'Artificial Intelligence: Past and Future' (2012), 55 (1) Communications of the ACM 5.

[78] McCarthy, J., 'President's Quarterly Message: AI Needs More Emphasis on Basic Research' (1983), 4(4) AI Magazine, 5

[79] Buchanan B.G., 'Artificial Intelligence as an Experimental Science' (1988) In: Fetzer J.H. (eds) Aspects of Artificial Intelligence. Studies in Cognitive Systems, vol 1. Springer, DordrechtHerbert A. Simon, 'Artificial Intelligence: An Empirical Science' (1995) 77, no. 2 Artificial Intelligence 95–127.

[80] Nick Bostrom, '*Superintelligence: Paths, Dangers, Strategies*' (2014) Oxford University Press, Inc. New York, NY, USA 32.

[81] Herbert A. Simon, (1995) 95–127

[82] Allen Newel and Herbert A. Simon, 'Computer Science as Empirical Inquiry: Symbols and Search, The 1976 ACM Turing Lecture' (1976) 19 Communications of ACM 113–126.

[83] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, et al., 'Artificial Intelligence and Life in 2030.' One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel' Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report.

[84] Nathalie Japkowicz, 'Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks' (2001) Vol 42, Issue 1–2, Machine Learning, 97–122

[85] High-Level Expert Group on Artificial Intelligence, A definition of AI: Main capabilities and scientific disciplines, 2019, p. 8.

essential.[86] The prohibition on solely automated decision-making in the GDPR does not pose a significant hurdle for AI research, just on the application of AI products and services. However, when AI products and services are implemented and impact citizen life, it is too late to protect them. In medical research, pharmaceutical companies need to follow strict rules and acquire permissions from the beginning of their research until their product is ultimately removed from the market. Even after it, they continue to be responsible for the effects of them. AI research requires similarly strict rules on research and development.

Public interest should be a crucial factor in supporting research activities.[87] For instance, when a company develops a chatbot for customer service, it does not have a clear public interest, and as a result, these kinds of AI products and services should not benefit from the GDPR research exemption. However, when the chatbot is developed for a hospital to communicate with patients, it may represent public interest in this particular healthcare context; thus, this type of research fits into the research exemption under the GDPR. The balancing becomes more challenging when the public interest is very much contextual-based, and the research is strongly attached to commercial activities, such as developing safer autonomous cars.

Interpreting the GDPR research exemption by private corporations to further use the data without public interest and to store it for indefinite periods might be considered abusive. Data subjects have several rights, such as the right to be forgotten, which could be derogated in the case of scientific research. Therefore, data protection authorities need to collaborate with national authorities responsible for oversight of scientific research to balance the data subjects' rights and the integrity of research. These measures aim to balance the lack of harmonised regulation of scientific research in the EU, and protect privacy without hindering innovation.

Conclusion

The research and development of AI products and services have a crucial impact on privacy. Therefore, the application of GDPR on AI research became a central issue for both public and private research organisations. AI research is a part of computer science; thus, the GDPR research exemption can be applied to it, allowing researchers to further use personal data. However, opaque algorithms, corporate secrecy and lower ethical standards in commercial research might pose a significant risk for citizens. Therefore, commercial AI research should not benefit from the GDPR research exemption without public interest and similar safeguards as academic research, such as review by independent ethics committees. Transparency and accountability can build trust together. In the case of AI research, governments and international organisations are expected to take more responsibility in exercising control over companies and privately funded research, which may counterbalance the lack of transparency and help citizens to build trust in the development of future technology.

Author contributions

Janos Meszaros and Chih-hsing Ho contributed equally to this article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by Academia Sinica's Academia Data Safety and Talent Cultivation Project and the Ministry of Science and Technology granted project on the Right to Explanation under EU GDPR? – An Exploration of the Due Process of Artificial Intelligence. (MOST: 109-2410-H-001-034, Taipei, Taiwan).

[86] See generally on regulatory issues about AI: Roger Clarke, 'Regulatory alternatives for AI' (2019) Volume 35, Issue 4 Computer Law & Security Review, 398-409

[87] See recommendations 16-23 of the Berlin Data Ethics Commission, Opinion of the Data Ethics Commission-Executive Summary, 22 October 2019. Available at https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.html;jsessionid=1B71C1E6D363C833EC7F485C2AF205AD.1_cid297?nn=11678512.