Note to Delegations:
This document is also available on O.N.E under the reference code:
DSTI/STP/NESTI(2019)1/FINAL

Acknowledgements

This report was prepared by Izumi Yamashita, Akiyoshi Murakami, Stephanie Cairns and Fernando Galindo-Rueda while they were based at the Science and Technology Policy (STP) Division in the OECD Directorate for Science, Technology and Innovation (DSTI). Brigitte van Beuzekom facilitated the use of bibliometric data to complement the analysis. This study has been part of the Programme of Work and Budget of the Committee for Scientific and Technological Policy (CSTP) and entrusted to the Working Party of National Experts on Science and Technology Indicators (NESTI).

The authors would like to express their gratitude towards Cecilia Cabello and Joseba Sanmartín (The Spanish Foundation for Science and Technology, Spain) and Jerónimo Arenas (University Carlos III of Madrid, Spain) for their support in analysing the Spanish and the European Commission CORDIS data; Alexandra Vennekens, Margot Schel, and Iris Glas (Rathenau Institute, Netherlands) for their support in analysing the Dutch data; Tamami Fukushi and Madoka Nakamura (both from Japan Agency for Medical Research and Development) for their support on providing the AMED data; and Hitoshi Koshiba (National Institute of Science and Technology Policy, Japan) for advice on text analysis and topic modelling. Voluntary contribution support by Japan's Ministry of Education, Culture, Sports, Science and Technology is also gratefully acknowledged.

Table of contents

Tables

Figures

Boxes

Abstract

This report presents the results of a proof of concept for a new analytical infrastructure ("Fundstat") for analysing government funding of research and development (R&D) at the project level, exploiting the wealth of text-based information about funded projects. Reflecting the growth in popularity of artificial intelligence (AI) and the OECD Council Recommendation on AI's emphasis on R&D investment, the report focuses on analysing government investments into AI-related R&D. Using text mining tools, it documents the creation of a list of key terms used to identify AI-related R&D projects contained in 13 funding databases from eight OECD countries and the EU, provides estimates for the total number and volume of government R&D funding, and characterises their AI funding portfolio. The methods and findings developed in this study, also serve as a prototype for a new expanded, distributed mechanism capable of measuring and analysing government R&D support across key priority areas and topics for the OECD and its member countries.

Keywords: Research and development, government funding, artificial intelligence

Synthèse

Ce document présente les résultats d'une preuve de concept pour une nouvelle infrastructure analytique (« Fundstat ») pour analyser le financement gouvernemental de la recherche et développement (R&D) au niveau des projets, en exploitant la richesse des informations textuelles sur les projets financés. Reflétant la popularité croissante de l'intelligence artificielle (IA) et l'accent mis par la Recommandation du Conseil de l'OCDE sur les investissements en R&D, il se concentre sur l'analyse des investissements gouvernementaux dans la R&D liée à l'IA. À l'aide d'outils d'exploration de texte, le rapport documente la création d'une liste de termes clés utilisés pour identifier les projets de R&D liés à l'IA contenus dans 13 bases de données de financement de huit pays de l'OCDE et de l'UE, fournit des estimations du nombre total et du volume de financement public de la R&D, et caractérise leur portefeuille de financement de l'IA. Les méthodes et les résultats développés dans cette étude servent également de prototype à un nouvel mécanisme étendu et distribué capable de mesurer et d'analyser le soutien gouvernemental à la R&D sur divers sujets prioritaires pour l'OCDE et ses pays membres.

Mots-clés : recherche et développement, financement public, intelligence artificielle

Executive summary

This document reports on the procedures and findings from an experimental text-based analysis of project-level research and development (R&D) funding data, focused on measuring the extent and features of government R&D support for Artificial Intelligence (AI-related R&D). The field of AI research has undergone a radical transformation in the past two decades, morphing from a small, relatively niche domain into a sprawling web of ground-breaking innovations. Mapping and measuring this research explosion – and the funding underlying its ignition – is of prime importance to policy makers and experts, as is encouraging its further development towards the common good. The 2019 OECD Council Recommendation on Artificial Intelligence states that governments "should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI". Tracking government investments into R&D is therefore of particular importance. While attempts have been made to assess government spending on AI-related R&D, mostly through proxy approaches, no comprehensive method exists by which to track and compare AI R&D funding across countries and agencies.

The study has used a quantitative case study approach, applying a set of text mining tools to specific project funding databases to identify AI-related R&D. The project-level funding data of 13 databases from eight OECD countries (Australia, Canada, France, Japan, the Netherlands, Spain, the United Kingdom and the United States) and the European Union provided useful and relevant ground for demonstration purposes. R&D project funding databases, while not indicative of total government R&D funding, can be used to trace and estimate a sizeable part of total government funding of AI-related R&D, thereby helping support implementation of the OECD Council Recommendation.

This study has adopted a "key terms" selection and matching approach for the identification of AI-related R&D projects by each organisation within the text corpus of their funded R&D projects. The task was to predict, using project titles and abstracts, whether or not a project was AI-related. Key term selection aimed to deliver a comprehensive list of AI-relevant key terms for document matching. A baseline set of potential key terms was enriched by means of text analysis applied first to a separate relevant body of scientific publication abstracts and then to the project funding text corpora that were the object of this study. A document was categorised as AI-related depending on the presence of key terms within it.

The total volume of AI-related government R&D funding identified through this exercise grew from USD 207 million in 2001 to almost USD 3.6 billion in 2019, a seventeen-fold increase. While this represents a very large amount, it may be dwarfed by business R&D investment if one considers that a single heavily AI-reliant company like Alphabet reported USD 20 billion worth of total R&D expenses in 2018, and the United States National Center for Science and Engineering Statistics provided a conservative estimate of business AI R&D investment in the order of USD 9 billion out of over USD 160 billion worth of total software R&D. Much of this surge in government funding is concentrated in recent years, with EU funding doubling in 2019. This pattern has had some ups and downs, however, reflecting the lumpiness of some large projects sponsored by R&D funding agencies and showing the importance of considering monetary measures as well as counts-based indicators.

When comparing sheer amounts of AI-related R&D funding by these agencies or bodies, the funding mechanisms covered by the EU's Community Research and Development Information Service (CORDIS) have very recently become the single largest source, followed closely by the US's National Institutes of Health (NIH) and its National Science Foundation (NSF). These two US agencies account for over three quarters of the cumulative AI R&D funding documented in this exercise. A different hierarchy emerges when examining the percentage of each agency's total R&D spending that is accorded to AI research, an indicator of AI R&D funding intensity. By this measure, the leading agencies are NSF, the UK Research Councils, the Dutch Research Council (NWO), and Innovate UK, each of which dedicated in 2019 between 10% and 15% of total R&D funding to AI-related projects.

The results clearly show that AI is not exclusively of interest to computer science and engineering funding agencies. To illustrate this point, the paper uses a statistical topic modelling technique to identify common topics (and group documents by these topics) on each database. The topics correspond to five broad themes: general AI techniques, AI prerequisites and impact (such as education and training and social impact), AI fields (such as computer vision and natural language processing), medical AI applications, and non-medical AI application areas (such as business and the social sciences). Around half of the funding streams for AI R&D projects focused most frequently on particular fields or techniques (e.g. computer vision), while the other half were principally concerned with particular applications of AI, either of a health/medical nature in the case of the three medical agencies covered and the funding streams incorporating support for business R&D and innovation.

This work also represents a pilot exercise in assessing the feasibility of constructing a multi-country infrastructure on R&D project funding for analytical purposes. Identifying emerging R&D domains and application areas is key in light of heightened interest in the directionality of R&D support by governments. The study has shown that decentralised, collaborative distributed approaches are possible and can deal with confidentiality considerations. Advances in data harmonisation are nonetheless necessary in order to enable future longitudinal and cross-country analysis. Different and evolving patterns of reporting and describing projects in application abstracts can lead to marked differences in results of text mining approaches. All these issues will be part of the remit of future OECD work on analysis of administrative STI financing data, in fulfilment of the OECD Blue Sky agenda for indicators.

AI is far from being the only research field that evades easy definitions but whose emergence remains critical to track. This pilot study will serve as a prototype for taking on the development of broader analysis mechanisms capable of assessing government contributions to a myriad of fields and applications, including pandemic resilience objectives and outcomes connected with the UN Sustainable Development Goals.

Résumé

Ce document présente les procédures et les résultats d'une analyse textuelle expérimentale des données relatives au financement de la recherche et du développement (R-D) au niveau projet, visant à mesurer l'ampleur et les caractéristiques du soutien public à la R-D en faveur du développement de l'intelligence artificielle (R-D liée à l'IA). Au cours des deux dernières décennies, le champ de la recherche en matière d'IA a connu une transformation radicale, passant d'un domaine mineur relativement spécialisé à un réseau tentaculaire d'innovations pionnières. Il est essentiel pour les décideurs et les experts de procéder à la cartographie et à la mesure de cet essor de la recherche - et du financement qui la sous-tend - tout comme d'encourager la poursuite de son développement dans l'intérêt commun. La Recommandation du Conseil de 2019 sur l'intelligence artificielle indique que les pouvoirs publics « devraient envisager des investissements publics à long terme et encourager les investissements privés dans la recherche et le développement, notamment interdisciplinaire, afin de stimuler l'innovation dans une IA digne de confiance ». Le suivi des investissements publics en faveur de la R-D revêt donc une importance particulière. Malgré des tentatives visant à évaluer les dépenses publiques de R-D liée à l'IA, principalement par le biais d'approches indirectes, il n'existe pas de méthode d'ensemble permettant de suivre et de comparer le financement de la R-D en matière d'IA dans différents pays et organismes.

L'étude s'est fondée sur une approche quantitative basée sur des études de cas, en appliquant un ensemble d'outils d'exploration de texte à des bases de données de financement de projets spécifiques afin d'identifier les activités de R-D liée à l'IA. Les données relatives au financement au niveau projet provenant de 13 bases de données dans huit pays de l'OCDE (Australie, Canada, Espagne, États-Unis, France, Japon, Pays-Bas et Royaume-Uni) et de l'Union européenne ont fourni des éléments utiles et pertinents pour la démonstration. Les bases de données relatives au financement de projets de R-D, bien que non révélatrices du financement public total de la R-D, peuvent être utilisées pour identifier et évaluer une partie importante du financement public total de la R-D en matière d'IA, contribuant ainsi à la mise en œuvre de la Recommandation du Conseil de l'OCDE.

Cette étude a recouru à une approche fondée sur la sélection et la mise en correspondance de « termes clés » pour l'identification des projets de R-D liée à l'IA soutenus par chaque organisation dans le corpus textuel de leurs projets de R-D financés. Il s'agissait de déterminer, à partir des titres et des résumés des projets, si un projet était lié ou non à l'IA. La sélection de termes clés avait pour but de dresser une liste globale de termes clés pertinents au regard de l'IA pour la mise en correspondance avec les documents. Nous avons enrichi un groupe de référence composé de termes clés potentiels au moyen d'une analyse de texte appliquée d'abord à un ensemble distinct de résumés de publications scientifiques dans le domaine concerné, puis aux corpus de textes relatifs au financement de projets qui ont fait l'objet de cette étude. Un document était qualifié comme étant lié à l'IA en fonction de la présence de termes clés en son sein.

Le volume total du financement public de la R-D liée à l'IA identifié dans le cadre de cet exercice est passé de 207 millions USD en 2001 à près de 3.6 milliards USD en 2019, ce qui représente une multiplication par dix-sept. Bien que cela représente un montant très important, il peut paraître minuscule par rapport aux investissements dans

la R-D des entreprises si l'on considère qu'une entreprise fortement tributaire de l'IA comme Alphabet a déclaré avoir dépensé à elle seule en 2018 un total de 20 milliards USD au titre de ses activités de R-D, et que le *National Center for Science and Engineering Statistics* aux États-Unis a fourni une estimation prudente des investissements des entreprises en matière de R-D liée à l'IA de l'ordre de 9 milliards USD sur un total de plus de 160 milliards USD en matière de R-D consacrée aux logiciels. Une grande partie de cette flambée des financements publics ses concentre sur les dernières années, le financement de l'UE ayant doublé en 2019. Cette tendance a toutefois connu des hauts et des bas, reflétant le caractère monolithique de certains grands projets parrainés par des organismes de financement de la R-D et montrant l'importance de prendre en compte les mesures monétaires ainsi que les indicateurs basés sur les dénombrements.

Si l'on compare le montant absolu du financement de la R-D liée à l'IA par ces agences ou organismes, très récemment les mécanismes de financement couverts par le service communautaire d'information sur la recherche et le développement (CORDIS) de l'UE sont devenus la source la plus importante, suivie de près par les *National Institutes of Health* (NIH) et la *National Science Foundation* (NSF) aux États-Unis. Ces deux agences américaines représentent plus des trois quarts du financement cumulé de R-D en matière d'IA documenté dans le cadre de cet exercice. Une hiérarchie différente apparaît lorsque l'on examine le pourcentage de dépenses totales de R-D accordées par chaque agence à la recherche en matière d'IA, qui représente un indicateur de la proportion de financement de la R-D dans ce domaine. Selon cette mesure, les agences principales sont la NSF, les conseils de recherche au Royaume-Uni, le conseil de recherche néerlandais (NWO) et *Innovate UK*, qui, en 2019, ont chacun consacré entre 10 % et 15 % du financement total de la R-D à des projets liés à l'IA.

Les résultats montrent clairement que l'IA n'intéresse pas exclusivement les organismes de financement dans les domaines de l'informatique et de l'ingénierie. Pour illustrer ce point, l'article utilise une technique de modélisation statistique des sujets afin d'identifier les sujets communs (et de regrouper les documents en fonction de ces sujets) au sein de chaque base de données. Les sujets correspondent à cinq grands thèmes : les techniques générales en matière d'IA, les conditions préalables et l'impact de l'IA (comme l'éducation et la formation et l'impact social), les domaines en matière d'IA (comme la vision par ordinateur et le traitement automatique du langage naturel), les applications médicales en matière d'IA et les domaines d'application non médicale de l'IA (comme le commerce et les sciences sociales). Environ la moitié des flux de financement de projets de R-D en matière d'IA se concentrent le plus souvent sur des domaines ou des techniques particuliers (par exemple, la vision par ordinateur), tandis que l'autre moitié s'intéresse principalement à des applications particulières de l'IA, notamment de nature sanitaire/médicale dans le cas des trois organismes médicaux couverts et des flux de financement intégrant le soutien à la R-D et à l'innovation des entreprises.

Ce travail constitue également un exercice pilote permettant d'évaluer la faisabilité de la construction d'une infrastructure multi-pays relative au financement de projets de R-D à des fins d'analyse. L'identification des domaines de R-D et des champs d'application émergents est essentielle compte tenu de l'intérêt accru pour la directivité à l'égard du soutien public à la R-D. L'étude montre que des approches distribuées décentralisées et collaboratives sont possibles et peuvent prendre en compte les aspects liés à la confidentialité. Des progrès en matière d'harmonisation des données sont néanmoins nécessaires afin de permettre la conduite de nouvelles analyses

longitudinales et transnationales. Des cadres différents et évolutifs pour la présentation et la description des projets dans les résumés des demandes peuvent entraîner des différences marquées au niveau des résultats des approches d'exploration de texte. Toutes ces questions seront abordées lors de travaux ultérieurs de l'OCDE sur l'analyse des données administratives relatives au financement de la STI, dans le cadre du programme *Blue Sky* de l'OCDE sur l'élaboration d'indicateurs.

L'IA est loin d'être le seul domaine de recherche qui ne se laisse pas aisément définir, mais il est essentiel d'en suivre les développements. Cette étude pilote servira de prototype en vue de l'élaboration de mécanismes d'analyse plus larges capables d'évaluer les contributions des pouvoirs publics à une myriade de domaines et d'applications, notamment les objectifs liés à la résilience aux pandémies et les résultats associés aux objectifs de développement durable des Nations unies.

Measuring the AI content of government funded R&D projects: A proof of concept for the OECD Fundstat initiative

1. Introduction and background

Artificial intelligence (AI) is transforming many aspects of our lives and influencing many decisions and processes. Rapid advances have resulted from research and development (R&D) efforts and from the widespread adoption of novel AI solutions, which in turn is generating expectations of further transformation and disruption to the way in which societies operate and address their current and future challenges. As reflected in the recent OECD Council Recommendation on Artificial Intelligence (OECD, 2019[1]), AI is a high priority in policy agendas[1] at both the national and international levels owing to its combined transformational and disruptive effects.

Government support for R&D has been central to the development of AI capabilities. According to the US National Research Council (NRC, 1999[2]), while the concept of AI originated in the private sector, its growth depended largely on public investments, from fundamental, long-term research into cognition to shorter-term efforts to develop operational systems. Leading government agencies included the Defense Advanced Research Projects Agency (DARPA), the National Institutes of Health (NIH), the National Science Foundation (NSF), and the National Aeronautics and Space Administration (NASA), which have pursued AI applications of particular relevance to their missions. From the 1960s through to the 1990s, DARPA provided the bulk of the United States' support for AI R&D through its Information Processing Techniques Office (IPTO) and thus helped to legitimise AI as an important field of inquiry while also influencing the scope of related research. According to Goldstein (Goldstein, 1992[3]), total federal funding for Artificial Intelligence research went from USD 105 million in 1984 to USD 274 million in 1988, with the share of "basic" research decreasing from 42 to 31 percent as interest in developing concrete applications surged.

More recently, a number of policy documents have drawn attention to the level of efforts made by governments to advance R&D on AI. The supplement to the US President's FY2021 budget on the Information Technology R&D Program indicated that the Federal non-defence budget for AI R&D reached USD 1.1 billion in 2019 and was expected to rise to USD 1.5 billion in 2021 (NSTC, 2020[4]). In the case of the People's Republic of China, Acharya and Arnold (Acharya and Arnold, 2019[5]) estimate using open sources that non-defence AI R&D spending was in the USD 1.7 to 5.7 billion range, i.e. of a comparable order of magnitude to that of the United States. The European Commission reported an increase of 70% in its annual investments in AI under the research and innovation programme Horizon 2020 and expected to reach EUR 1.5 billion (ca. USD 1.8 billion) over the 2018-2020 period (EC, 2018[6]). In the United Kingdom, the Engineering and Physical Sciences Research Council (EPSRC) is reported to have allocated GBP 300 million (ca. USD 410 million) to fund research

[1] See further examples at http://www.oecd.org/going-digital/ai/initiatives-worldwide/

related to data science and AI (BEIS and DCMS, 2018[7]). Within the European Union, the French government set out the intention to invest EUR 100 million in AI research projects through its national research funding agency (ANR) between 2018 and 2022 as one of the pillars of its AI strategy (MESRI and DINUM, 2018[8]), while in Spain, state bodies were estimated to be contributing EUR 114 million to R&D and related activities in this area (MICINN, 2019[9]). The Australian Research Council has awarded over AUD 243 million on pure research projects formally classified as AI and image processing since 2010 (Hajkowicz SA et al., 2019[10]).

In this context, the 2019 OECD Council Recommendation explicitly states that governments "should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI […]". This specific element of the recommendation lends itself to monitoring across countries and over time. However, no effective benchmarks exist for total nor for government-funded R&D on AI, especially for enabling international comparisons as is commonly aimed for. Initiatives like Stanford's AI Index (Perrault et al., 2019[11]) assess the state of AI R&D activities through proxy measures based on scientific publication, conference, and patent data, without actually providing information on the value of public or private efforts to advance the state of the art in AI.

Emergent and rapidly evolving domains of enquiry or technology present multiple measurement challenges, which preclude the existence of readily available and tested statistical indicators at the desired level of granularity. In the case of classifications for R&D expenditures and funding, the international consensus reflected in the OECD *Frascati Manual* (OECD, 2015[12]) is limited to a high-level classification system, which provides only indicative guidance for the regular annual national reporting of statistics, within which AI as a field would be part of the *Computer and Information Sciences* subdomain of *Natural Sciences*. Official statistics take time to reflect emergent domains in the field. Definitions must be developed or adapted, and chains of resources and activities, such as surveys, must be put into motion, ultimately resulting in new statistical indicators that meet the needs of different types of user. The coordination of such exercises at the international level can be jeopardised by the expectation that, by the time such a production cycle has been completed, R&D domains may have evolved, and priority questions changed. In practice, some countries are keener to adopt a standard of their own choosing, while others will prefer to wait before committing to a particular approach.[2] From the experience of biotechnology and nanotechnology R&D statistics, countries appear to paradoxically find it easier to report measures of business expenditure than government expenditure. It might appear that this is due to the fact that individual government agencies may themselves lack the administrative data infrastructure and mandate that allows them

[2] A number of countries have experimented with AI related questions in their business R&D and innovation surveys, using different definitions. There is at present no international consensus on how such data should be collected. At the 2019 NESTI workshop on innovation surveys and the implementation of the 2018 *Oslo Manual*, participants consistently highlighted this topic as a high priority for addressing recommendations on measuring "digital innovation". Different examples for business ICT surveys have been documented in OECD (2021), "AI Measurement In ICT Usage Surveys: A Review", OECD Digital Economy Papers, No. 308, OECD Publishing, Paris, https://doi.org/10.1787/20716826

to submit a figure that may have additional administrative implications, and NSOs may lack the powers to compel them to provide a figure.

Faced with this challenge, which is not exclusive to AI but is present in any other emergent R&D domain, the OECD *Blue Sky Forum* on the future of science and innovation data and indicators held in 2016 posited the possibility of developing alternative, complementary pathways to the statistical analysis of R&D funding. The notion of an "R&D project" as unit of analysis was explicitly introduced in the 2015 edition of the *Frascati Manual* with the aim of facilitating data generation efforts that underpin the production of R&D statistics. Data about R&D projects can be extremely rich sources of information, as they allow information reporters to establish whether these projects truly are R&D projects. Additional features that enable tagging and classification can also be extracted from such data. The proposals at the *OECD Blue Sky Forum* (OECD, 2018[13]) went one step further by proposing the active and coordinated use of data about R&D projects, under what came to be described as the ***Fundstat* infrastructure** concept.

The basis of *Fundstat* as a potential statistical infrastructure is that project text descriptions across several funding agencies can be readily analysed in combination with funding amounts and other connected data with the help of text mining techniques to implement flexible and adaptable indicators. For this idea to result in a well-functioning data and analysis infrastructure, it is also necessary to put in place the mechanisms across funding agencies and their data resources that enable the on-demand, consistent implementation of data queries that generate outputs at the desired level of conceptual granularity and that complement existing indicators. Enquiries into the rate of public investments in R&D for AI appear to lend themselves to the *Fundstat* approach because of the growing pervasiveness of potential AI applications and the lack of a definitive and rapidly implementable definition for "AI R&D" that can be tested against different possible uses.

Administrative data on R&D project funding by governments are complex objects of statistical analysis. They are not systematically available across countries and agencies, very different formats are used, and their coverage represents a variable and not necessarily representative part of all government R&D funding. The project level data are only representative of total R&D funding depending on the extent to which project-based funding is the norm within a country. In systems where institutional block funding is the major resource allocation mechanism it is more difficult to claim a certain degree of representativeness, but the data are informative of the government's discretionary use of R&D funds and other non-discretionary instruments where there may be also be project-level data available for analysis.

Intrinsic data quality is a major consideration in deciding whether project funding databases are suitable for statistical analysis. Because of the administrative nature of the data, specific purposes of both funders and applicants underpin the data generation process and constrain the range of admissible interpretation of indicators based on such data. In the case of funded R&D projects, the depth and breadth of the information provided by applicants will depend on the incentives and constraints that they face. Tagging of projects by applicants or administrators is also potentially subject to human error and inconsistent applications of definitions, a problem that can impact both designed and data-driven approaches to indicators.

This document reports on the procedures and initial findings from an experimental text-based analysis of project-level R&D funding data, focused on measuring the

extent and features of government support for projects with AI content (AI-related R&D). The study applied a set of quantitative tools to identify **AI-related R&D** in project funding databases, contributing to the *Fundstat* proof of concept. The project-level data on R&D funding from 13 databases covering funding agencies and programmes in eight countries and the European Commission provided useful and relevant ground for conducting this exercise across different countries and institutional settings for demonstration purposes.

In order to aid with the monitoring of how OECD countries and partner economies invest in AI R&D, this study's primary operational objective was to assess whether and how it is possible to identify AI-related R&D projects, namely **R&D projects whose text descriptions render themselves suitable to be classified as seeking advances in AI or as making an explicit and non-trivial use of AI systems to achieve their objectives**. Based on the outcome of the (approximate) identification of the full set of AI-related projects in a selected corpus of project funding databases, the statistical goal was to estimate the volume and share of projects and funding amounts that fit into this category.

The ultimate purpose of this study is twofold:

- To inform policy discussions on public support for AI-related R&D (thereby contributing to wider OECD efforts in this area[3]) by increasing understanding of the transformational role of AI as a general-purpose technology that can also enable R&D and innovation in different scientific domains and application areas.
- To support the OECD pilot assessing the feasibility of constructing a multi-country infrastructure on R&D project funding for analytical purposes ("Fundstat"), which would include the identification of emerging R&D domains and application areas, in light of heightened interest in the directionality of R&D support by governments (OECD, 2021[14]). There is particularly high interest in measuring the contribution of government R&D funding to narrowly defined Sustainable Development Goals (SDGs) or pandemic resilience.

Therefore, beyond the concrete application to AI as the area subject to exploration, this work seeks to address the widespread demand for data resources, tools, and methods that help identify features of R&D funding in thematic areas that are not easily captured by pre-defined and difficult-to-change taxonomies. This exercise is furthermore a demonstration of the possibilities of AI methods for text analysis as complementary detection and classification tools for statistical measurement that enable greater analysis uniformity and replicability. The increasing public availability of project-level funding data, often due to public transparency measures, is also enabling related efforts looking specifically at data about R&D funding. Funding organisations and a growing number of commercial providers of research support services have been compiling and offering access to data and providing semantic search and analytical functionalities (Bode et al., 2019[15]).

This study provides a unique perspective to the analysis of project-level data as it combines databases from different countries using a mixed approach that combines

[3] See https://www.oecd.org/going-digital/ai/

centralised pooling[4] of publicly available sources with a federated approach, which does not require OECD to have direct access to the project-level data. This therefore represents an initial demonstration of a model by which participating agencies can agree on protocols for distributed data analysis and for the potential exchange of information subject to predefined disclosure rules. Under a universal fully open access model, OECD's role in this space would be limited, as research groups would be well equipped to analyse such data. However, while there is a widespread shift towards increased openness of R&D funding project microdata, it has become clear that not all agencies are willing to place all potentially relevant data in the public domain. In light of the growing need for the coordination of data exchanges and analysis, the OECD is well positioned to assist in this role, as its experience with national statistical agencies and confidential business survey data through the microBeRD project (OECD, 2020[16]) shows that distributed analysis mechanisms represent a feasible second-best solution when data conform to basic minimum common standards.

This document is structured as follows. **Section 2** describes the data used for analysis and the methodology applied to identify AI-related R&D funding in the absence of a proper training database in which projects have been comprehensively *ex-ante* rated as AI-related. **Section 3** presents the key results for the 13 databases from eight countries and a region, reporting fast growing levels and rates of AI-related funding, and providing additional evidence on the topics of the projects identified as recipients of AI-related R&D funding. Complementing these, agency-specific results and robustness checks are also available in the Annex. **Section 4** concludes by outlining the planned next steps for the work on project level R&D funding data.

2. AI-related project retrieval methodology

2.1. Project funding data

There is a wide and fast growing literature dealing with field-specific topic extraction from several corpora, mostly publications, with some efforts looking at AI in particular, as documented in (Cockburn et al., 2018[17]) and previous OECD work aimed at identifying and measuring Artificial Intelligence (AI)-related developments in science, as captured in scientific publications; technological developments, as proxied by patents; and software, particularly open source software (Baruffaldi et al., 2020[18]). However, there are fewer precedents when it comes to R&D funding data (Abadi, He and Pecht, 2020[19]; Annapureddy et al., 2020[20])[5]. Project funding data are products of the administrative processes of R&D funding organisations (ministries, agencies, etc.), performed either internally or externally, to select, fund, and monitor R&D projects. The data used in this study came from 13 organisations deemed to be primarily concerned with R&D funding activities.

[4] For example, the best-known data source on R&D grant data, the Dimensions database by Digital Science, incorporates a pool of grant data accessible to Digital Science but does not include within its coverage projects from agencies that do not wish to share this type of data.

[5] A number of papers investigated R&D funding data dedicated to health sciences: (Schmutz et al., 2019[30]), (Scarpelli, Whelan and Farahani, 2020[31]) and (Gallo et al., 2020[32]).

The 13 R&D and innovation funding databases from authorities or agencies in eight countries and the EU used in this paper are as follows:

- The Australian Research Council (ARC).
- The Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council (NSERC).
- The programmes under the Spanish National Plan for Scientific and Technological Research and Innovation (PlanEst), covering multiple state-level bodies.
- The French National Research Agency (ANR).
- The United Kingdom's Gateway to Research (GtR), which contains data for the seven research councils (GtR_RC) and the Innovate UK (GtR_Inno) [6].
- Japan's Agency for Medical Research and Development (AMED) and Database of Grants-in-Aid for Scientific Research (KAKEN).
- The Dutch Research Council (NWO).
- The United States' National Institutes of Health (NIH) and National Science Foundation (NSF).
- The European Commission's Funding Programmes covered by the Community Research and Development Information Service (CORDIS).

When considering the data, it is important to have a basic understanding of the missions and types of programmes of the funding organisations.

- With some exceptions, most of the organisations operate primarily in the basic and applied research space of the R&D spectrum, fostering advances in fundamental knowledge and research into potential applications while refraining from funding the experimental development of products or processes for commercialisation. A number of projects may include activities that do not fully qualify as R&D, such as other types of S&T or innovation activities. For the selected agencies, this may be particularly the case of S&T infrastructure projects, but as such projects are primarily intended to contribute to R&D activities, no attempt was made to identify and remove them.
- All of the organisations provide financial support in the form of grants, cooperative agreements, and contracts, making extensive use of peer review as a resource allocation mechanism. Other instruments such as loans are significantly less common and important in volume in the majority of cases.
- As previously noted, not all the databases are openly available for download and analysis. This is the case of the Dutch and Spanish databases. AMED data are open but not available for download.[7]

[6] The analysis separates these two groups that are now integrated under UK Research and Innovation.

[7] Japan's AMED data is open, but downloadable. The authors asked the agency for their data provision.

- Project descriptions are obtained by authorities through the funding application processes. Only awarded projects are available and were included in the study.[8]

The analysis used all data available from between 2001 and 2019 in each database. Basic information on the number of projects[9] and the funding volumes for each database are available in **Table 2.1**. The biggest database in terms of total amount of funding is the NIH, followed by the CORDIS and the NSF.

Table 2.1. Main features of the databases analysed

Database	Countries/Region	Available period	Number of projects	Total amount of funding (USD Million)	Language	Data access	Analysis approach
ARC	Australia	2002-2019	26 677	8 994	English	Open	Pooled OECD
CIHR	Canada	2001-2018	56 778	14 147	English or French	Open	Pooled OECD
NSERC	Canada	2001-2017	175 945	3 402	English or French	Open	Pooled OECD
PlanEst	Spain	2004-2016	67 770	22 256	Spanish	Confidential	Distributed
ANR	France	2005-2019	20 123	6 506	French	Open	Pooled OECD
GtR_Inno	United Kingdom	2008-2019	18 424	14 281	English	Open	Pooled OECD
GtR_RC	United Kingdom	2006-2019	80 736	46 280	English	Open	Pooled OECD
AMED	Japan	2015-2018	4 765	4 213	Japanese	Open	Pooled OECD
KAKEN	Japan	2001-2018	466 709	33 750	Japanese or English	Open	Pooled OECD
NWO	Netherlands	2016-2019	7 177	2 186	English or Dutch	Confidential	Distributed
NIH	United States	2001-2019	1 428 472	497 955	English	Open	Pooled OECD
NSF	United States	2001-2019	224 307	114 883	English	Open	Pooled OECD
CORDIS	European Union	2001-2019	72 061	142 864	English	Open	Distributed

Note: The number of projects is based on decisions to allocate funding to the projects. In the case of the NIH, a project may be financed multiple times with multiple decisions, which are counted as multiple projects in the analysis.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

The databases under study offer relevant insights on the methodology of project description analysis and the role of AI in research. In the case of the United States, for example, NIH funding accounted for approximately 27% of total US federal government for R&D (obligations) in 2018, while NSF funding accounted for 4.8 % (44% and 8%, respectively, as a proportion of non-defence federal R&D obligations) (National Science Foundation, 2020[21]). In contrast, in the case of France, where 35% of all government R&D budgets are provided on the basis of general university funds[10], ANR funding accounted for only approximately 5% of government R&D funds in 2017 (MESRI, 2020[22]). Partial coverage of R&D programmes within a country also needs to be considered, in order to illustrate what type of R&D funding is excluded.

[8] Rejected proposals are not included unless they turned out to be successful at receiving funding at a late stage. The comparison of selected and rejected proposals can also be highly informative for a range of policy analysis purposes.

[9] This is based on the decision to allocate funding to a project. In the case of the NIH, a project may be financed multiple times through multiple decisions: this is counted as multiple projects in the analysis.

[10] These are government funds not earmarked for R&D that are discretionarily used by universities for R&D.

The omission of agencies at different government levels and programmes can result in systematic bias if attempting to extrapolate to the country as a whole.

Funding data typically contain the following attributes.

- Project ID: an identification number given to each project by the funding organisations during their administrative process.
- Project title and abstract: text descriptions of the project that typically contain the purpose, methodology, and expected outcomes of the project, which are used to classify the projects through text analysis.
- Date of award and project period.
- Funding amount: the amount of money allocated to the project.
- Information on the main beneficiary or principal investigator (PI) (e.g. PI ID, PI name, and beneficiary organisation): information concerning the people and organisations that obtained the funding allocation. These have not been used in the analysis.
- Keywords: words written by the applicants for the funding or by the funding organisation that represent the nature of the project. These have not been used in the analysis.
- Programme information: the administrative mechanisms through which the funding was allocated.
- Idiosyncratic thematic classifications: a category based on the scientific discipline of the project, which is often specified by the funding organisation using their own definitions. These have not been used in the analysis as they are not consistent between organisations nor granular enough to capture inter-disciplinary technologies such as AI.

In some instances, databases also include information on project outputs collected during and after the project. These have not been used in the current analysis.

Because of the lack of systematic tagging information within the 13 databases that is relevant to the objectives of this study, the analysis focused on text information included within project titles and abstracts[11]. As it does not depend on tagging features specific to any given database, this approach is relevant for potential application to other project funding databases.

2.2. Operational definition of AI

Definitions guiding statistical measurement work vary, and their operationalisations likewise vary according to the type of data used. Because this exercise uses existing administrative data for a secondary identification/classification purpose that differs from the data's original agency-specific grant management purpose, the role of an AI

[11] Although some of the data contain "project terms" that represent the main topics of the project, this analysis does not rely on them in order to maintain a common methodology for the funding agencies.

definition in this context is purely aimed at helping ensure the consistency of the selection approach (the procedures used and their outcomes).[12,13]

As noted in the *Frascati Manual*, R&D projects and the resources invested in them can be classified into a given field on the basis of project content similarity. This is a multidimensional concept that includes:

- The objects of interest – the phenomena to be understood or the problems to be solved through R&D.
- The knowledge sources drawn upon for the R&D activity to be carried out. Two projects are related if they share prior literature that they cite as relevant.
- The methods, techniques, and professional profiles of the scientists and other R&D workers – domains are often classified on the basis of the methodological approaches to the study of a given phenomenon or question.
- The areas of potential or envisaged application of the project results, particularly in the case of applied research and experimental development.

The analysis of taxonomic systems used by the 13 funding organisations revealed the existence of a domain-based definition for Artificial Intelligence as the "*theory and development of computer systems which perform tasks that normally require human intelligence. Such tasks may include speech recognition, learning; visual perception; mathematical computing; reasoning, problem solving, decision-making, and translation of language* (NIH MeSH)". This definition was extracted from the NIH MeSH (Medical Subject Headings), a hierarchically organised set of keywords managed by one of the NIH institutes (U.S. National Library of Medicine). The definition includes a list of potential application tasks and makes explicit reference to the concept of intelligence without defining it. The reference to "normally require HI (human intelligence)" is indicative of the potential subjectivity and context-dependence of the concept. Over time and with growing levels of automation, a number of tasks will ultimately cease to be considered AI depending on who makes the judgement.

The OECD Advisory Expert Group on Artificial Intelligence (AIGO) has defined AI not as a standalone concept but by reference to **AI systems**, namely as *"machine-based systems that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments […]by using machine and/or human-based inputs to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models through analysis in an automated manner (e.g. with machine learning, or manually); and iii) use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy"*. (OECD, 2019[23]).

This definition may be connected to the OECD definition of R&D in multiple ways, as AI systems may be instruments, objects, or intended outcomes of R&D projects.

[12] A designed survey-based approach, for example, would use a definition as a starting point for developing and testing question items aimed at implementing the chosen definition. A definition could also be offered to survey respondents to inform and support their own information retrieval and response processes.

[13] Ultimately, measurement and administrative purposes might align, although this may be realised too late or occur over a limited period of time.

R&D comprises creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture, and society – and to devise new applications of available knowledge (OECD, 2015[12]). Comprising basic and applied research as well as experimental development, all R&D projects are required to meet a set of criteria (creativity, novelty, uncertainty, systematicity, and transferability) in order to be considered as such. A hard definition of AI R&D could be thus conceived as representing R&D pursuing advances in the state of the art in knowledge about AI systems and their performance, while a more inclusive definition would include R&D projects where the use or development of AI systems may not be the primary motivation but would instead play an instrumental role.

Both the MeSH and OECD-AIGO definitions of AI underpin the work in this report, which aims not to propose new definitions of AI but rather to identify funded R&D projects that relate to AI in the sense that they either make use of AI systems or contribute to their development theoretically or practically. This is what is implied by the rather broad notion of "relatedness". This approach ultimately encapsulates the idea of data-driven implementations of AI R&D definitions, in contrast to designed approaches where the task of implementing a "designed" and tested definition is delegated to the reporting sources who are deemed to be in a position to judge. Further refinements in the methodology of this study might ultimately allow for more reliable differentiations between different types and grades of relatedness, particularly differentiation between research that makes use of available AI systems and tools (instrumental relatedness) and research that seeks to develop new AI concepts, theories, and tools (output relatedness), as well as possible instances when both are combined.

2.3. AI-related project retrieval methodology

The methods for AI-related project retrieval used in this study were initially developed for the NIH[14] and NSF databases and subsequently applied to the additional databases incorporated in the analysis.

After considering several options, a "key term" (sometimes referred to as keyword for simplicity) selection and tagging approach was adopted to identify AI-related R&D projects in the text corpus of funded projects for each organisation. The fundamental task was to predict a category using text data in the absence of consistently labelled data – that is, in the absence of R&D project records tagged as AI-related.[15]

[14] Some studies have looked at the classification of NIH funding ((Park et al., 2016[33]) and (Talley et al., 2011[28])), and of NSF funding ((Kawamura et al., 2018[27]) and (Freyman, Byrnes and Alexander, 2016[29])), but none have specifically targeted AI-related research.

[15] The embedded tagging of available records could be an important source of information. However, the thematic classification item in the NIH data (NIH spending categories), which is based on the Research, Condition, and Disease Categorization Process (RCDC), is not on its own an appropriate basis for identifying AI-related research because it is health objective-oriented and not informative about actual research methods. AI-related terms do not appear to be used comprehensively and tagging has not been consistent over time, as projects were manually tagged from 2001 to 2007 and automatically from 2008 to 2016. Furthermore, some documents do not include project description terms. In the NSF data, thematic classification items and project terms are not available. Similar situations apply to other databases analysed.

The potential use of fully unsupervised topic modelling – the identification of hidden semantic structures - in the full corpus of projects was discarded after an initial attempt. An examination of the results of conventional topic modelling schemes in the full R&D project corpus showed that topics proved difficult to associate with the notion of AI-relatedness as the AI signal was relatively weak, especially in the case of data from funding organisations dedicated to life sciences (i.e. CIHR, Canada; AMED, Japan; and the NIH, USA). In those databases, topics are often dominated by the semantic weight of the R&D objectives that are of importance to funders (for example, health outcomes) over the scarcer information about potential AI-related methods used in the research. Topic identification from internal or external linkages to other data (e.g. citations) was not possible either. With further refinements, it may ultimately be possible to revert to this type of modelling. For the time being, a somewhat simpler key term selection and tagging procedure has been followed.

Key term matching presents a number of challenges, since there is at present no consensus on a standard set of key terms that comprehensively and unambiguously represent AI-related R&D; moreover, such a set is bound to be specific to different *corpora* (scientific publications, R&D project proposals, patent claims, job descriptions, company reports, etc…) and vary over time. As noted in Baruffaldi et al. (2020), failing to capture all potentially relevant key terms risks overlooking many AI-related projects, thus underestimating their total number. This problem can also arise when the title and abstract of project applications do not contain sufficient information on the research methodology to be used in a given project. This is an underlying data problem, which may be particularly acute when AI plays an enabling role within a project (the notion of instrumental relatedness alluded to above), but the project's abstract focuses on outlining the expected outcomes. Ideally, the underlying project text corpus available for analysis should contain a "methods" section to facilitate a more effective data mining process as well as a better understanding of the role played by AI in various R&D fields.

Conversely, effective key term matching from a predefined menu of AI terms does not ensure that the research is ultimately AI-related. As potential key AI terms may describe research paradigms or domains that do not necessary relate to AI, it is possible to overstate the true volume of AI-related projects. It is quite common to employ terms from different scientific and technological domains as metaphors to allude to newly discovered concepts. An often-cited example is the term "neural network" that was "borrowed" by computer scientists from neuroscience. In addition, the use of terms relating to statistical analysis tools known for several decades and recently popularised and adapted for AI applications (e.g. Markov decision processes) may result in non AI-related projects being mistakenly classified as AI-related.

The study involved two main steps, namely 1) key term selection and categorisation on the basis of AI-relatedness, and 2) document classification based on an AI-relatedness selection rule. These steps were complemented by additional bias and robustness checks, followed by an exploratory topic analysis of the projects subsequently identified as AI related. These are summarised in **Box 2.1**:

Box 2.1. Sequence of steps for identifying AI R&D projects

1. Identification of key AI terms

1.1. Create a "base" list of key AI terms.

- The base list combines two existing lists, one extracted from the NIH Medical Subject Headings (a hierarchically organised set of keywords) and one produced by Cockburn et al. (2018). We refer to these two existing lists as the M-list and C-list, respectively.

1.2 Enrich / extend list of AI terms through identification of similar terms.

- A word-embedding similarity approach is employed.
  - Each word in a target database is vectorised – that is, transformed into a numerical vector or "word embedding" – using a standard word embedding model (word2vec). The base AI terms are also vectorised.
  - The vectorised words are then compared. If a word in the database is highly similar to the *base* AI terms, as measured by a mathematical notion of similarity, the word is retained as a possible candidate for the extended list of key AI terms.
  - Of these potential candidates, words whose connection to AI is deemed to be overly ambiguous are rejected, while all other words are added to the extended list of terms.
- Databases are employed in sequence for this enrichment process.
  - First, utilising the base list, additional key AI terms are extracted from a corpus of scientific publications in computer science journals classified as AI journals by publishers.
  - Second, utilising the newly extended list, additional key AI terms are extracted from all of the R&D project funding databases (multiple agencies).

1.3 Categorise all key AI terms as "core" or "non-core".

- Core-terms are deemed to imply no ambiguity.
- Non-core terms are somewhat ambiguously related to AI; for example, "neural network" is deemed non-core as it may refer to either an artificial or a biological neural network.

2. Classification of documents as AI related

2.1. Adopt a rule to classify documents as AI-related or non AI-related. A document was selected as (likely to be) AI-related if

- At least one core key term was found within its title or abstract; or
- Two or more distinct non-core terms were found (a special rule is applied to the terms "bioinformatics" and "computational biology"; in that case, an additional key term is required, as that particular is likely to produce false positives)

2.2. Classify documents according to rule.

2.3. Carry out a number of manual robustness checks in order to identify the degree of possible error.

3. Topic modelling of AI-related projects

3.1. For each funding database, use Latent Dirichlet Allocation (LDA) to generate nine or twelve topics (depending on the database), each of which are associated with words found in the database. LDA then maps each document in the database to one of those topics.

3.2. Manually label the topic subjects based on an examination and interpretation of the terms present within each topic.

2.3.1. Selecting key AI terms

Selecting key terms is itself a multi-step process. As there is no formal training database, the process requires an informative initial (base) list of terms that can be posited to be AI-relevant. This base list can then be extended into a longer list of AI-related terms that can be categorized as "core" and "non-core".

Step 1: Base AI terms

As previously noted, the MeSH taxonomy contains a heading for AI. **Table 2.2** provides a description of the position of AI within its hierarchical structure. AI features in two separate MeSH domains: Mathematical Concepts and Information Science. Subject subheadings include Biological Ontologies, Computer Heuristics, Expert Systems, Fuzzy Logic, Gene Ontology, Knowledge Bases, Machine Learning (including Supervised Machine Learning and Unsupervised Machine Learning), Natural Language Processing, Neural Networks (Computer), Robotics, and Support Vector Machines. This structure does not in and of itself provide a comprehensive source of all potentially relevant AI terms but instead provides a basic structure for the categorisation of research activity and outputs in the health domain. Unfortunately, the MeSH research tagging system is not yet applied either manually or automatically to funding applications. MeSH is principally applied to scholarly publications listed in MEDLINE, PuBMed, and related sources, while the funding data do not contain readily available information on publication outputs associated with the funded projects.

Table 2.2. MeSH tree structure for Artificial Intelligence

Mathematical Concepts [G17]
Algorithms [G17.035]
Artificial Intelligence [G17.035.250]
Machine Learning [G17.035.250.500]
Supervised Machine Learning [G17.035.250.500.500]
Support Vector Machine [G17.035.250.500.500.500]
Unsupervised Machine Learning [G17.035.250.500.750]
Information Science [L01]
Computing Methodologies [L01.224]
Algorithms [L01.224.050]
Artificial Intelligence [L01.224.050.375]
Computer Heuristics [L01.224.050.375.095]
Expert Systems [L01.224.050.375.190]
Fuzzy Logic [L01.224.050.375.250]

Knowledge Bases [L01.224.050.375.480]
Biological Ontologies [L01.224.050.375.480.500]
Gene Ontology [L01.224.050.375.480.500.500]
Machine Learning [L01.224.050.375.530]
Supervised Machine Learning [L01.224.050.375.530.500]
Support Vector Machine [L01.224.050.375.530.500.500]
Unsupervised Machine Learning [L01.224.050.375.530.750]
Natural Language Processing [L01.224.050.375.580]
Neural Networks (Computer) [L01.224.050.375.605]
Robotics [L01.224.050.375.630]

Source: U.S. National Library of Medicine, NIH. Extracted on 28 September 2018 from https://meshb.nlm.nih.gov/record/ui?ui=D001185.

Table 2.3. AI term list from Cockburn et al. (2018)

Symbols	Learning	Robotics
natural language processing	machine learning	computer vision
image grammars	neural networks	robot
pattern recognition	reinforcement learning	robots
image matching	logic theorist	robot systems
symbolic reasoning	bayesian belief networks	robotics
symbolic error analysis	unsupervised learning	robotic
pattern analysis	deep learning	collaborative systems
symbolic processing	knowledge representation and reasoning	humanoid robotics
physical symbol system	crowdsourcing and human computation	sensor network
natural languages	neuromorphic computing	sensor networks
image alignment	decision making	sensor data fusion
optimal search	machine intelligence	systems and control theory
	neural network	layered control systems

Source: Cockburn et al. (2018), The Impact of Artificial Intelligence on Innovation, http://www.nber.org/papers/w24449.

For this reason, the MeSH list of terms (the "M-list" from here onwards) was enhanced with a key AI term list produced by (Cockburn et al., 2018[17]), who analyse academic papers and patent documents to measure the impact of AI on innovation and derive a list of key terms related to AI as a basis for their analysis. This list ("C-list") contains 38 terms classified into three categories (Symbols, Learning, and Robotics) and is reproduced in **Table 2.3**. The C-list is more comprehensive than the M-list, although the former contains terms that are not uniquely associated with AI and may therefore imply a lower degree of precision.

Step 2: Extending the set of potential key terms by analysing scientific publication data

Combined into one, the M- and C-lists represent a possible a base set of key terms.[16] However, before proceeding further, additional work was required to minimise the

[16] For further analysis to be described below, the terms in the combined M-C list were converted to lower case text and lemmatised, i.e. variations of the same terms were converted to a single item (e.g. sees, saw, seeing, seen to see, or books to book). The terms "supervised machine learning" and "unsupervised machine learning" were converted to "supervised

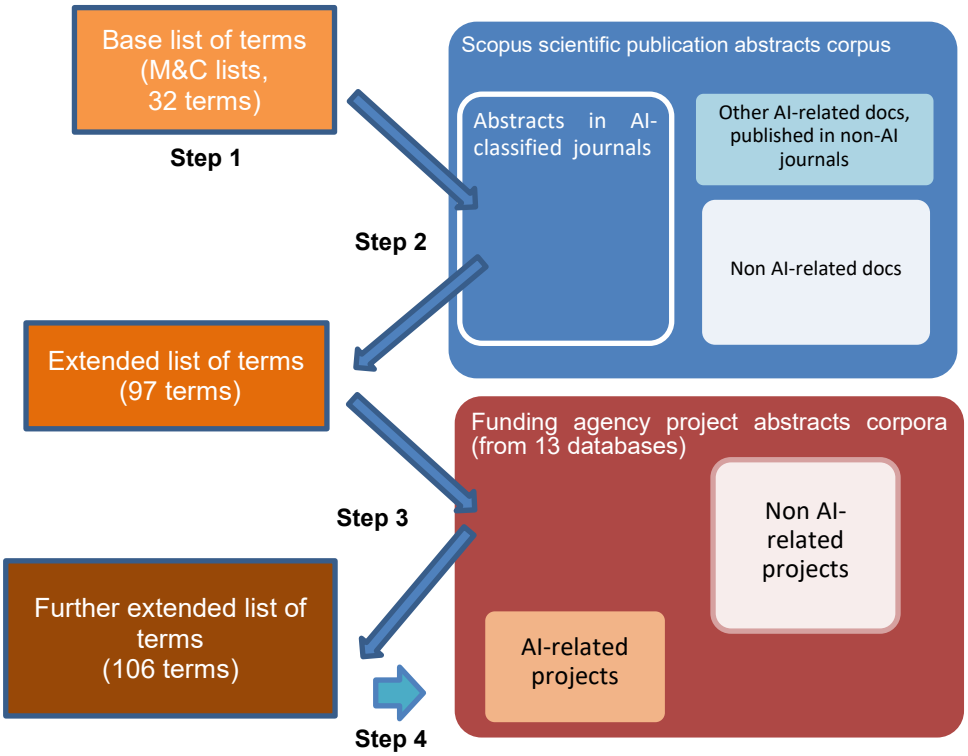
risks of high false discovery or omission rates. In particular, a process was required to retrieve additional key terms that provide relevant signals of AI-related research activity and that minimise the false omission rate. This procedure is outlined in **Figure 2.1**.

A reasonable data corpus from which to retrieve an extended set of baseline terms is the body of scientific publications (articles, conference proceedings, and reviews) featured in journals and dissemination vehicles known to focus on AI. The Scopus Custom database used at OECD provides titles and abstracts for 713 016 documents published between 2001 and 2017 that have been assigned the All Science Journal Classification (ASJC) codes corresponding to Artificial Intelligence (ASJC1702) and Computer vision and pattern recognition (ASJC1707).[17]

learning" and "unsupervised learning" as they overlap with "machine learning". The terms "decision making" and "natural languages" were removed from the core combined M-C list as they were deemed potentially ambiguous. Furthermore, in order to reduce the level of noise in the word embedding process, a number of terms with very low incidence in all the corpora used for text analysis were removed, namely "biological ontologies", "computer heuristics", "crowdsourcing and human computation", "image grammars", "layered control systems", "logic theorist", "physical symbol system", "symbolic error analysis", and "symbolic processing". Thus, a shortlist of 32 candidates of Key AI terms was retained.

[17] The Scopus database does not, however, provide a full basis for training because there is no built-in identification of AI documents outside the corpus of documents published in AI-classified journals.

Figure 2.1. Outline of Key AI term identification and tagging procedure



Step 1: Obtain base list of terms from previous studies
Step 2: Extend base list of terms by analysing Scopus (only AI-classified journals)
Step 3: Further extend the extended list of terms by analysing the 13 funding databases
Step 4: Label key terms as "core" and "non-core" to tag and classify AI-related projects
Robustness check of the tagging with selected funding databases (Annex D)

Note: This figure provides a schematic representation of the procedure, which begins with the base list of terms extracted from the M- and C-lists. This is followed by the retrieval of terms used in similar contexts within the corpus of scientific publications in AI journals and within the corpus of funding agency documents. The procedure concludes with the applications of the definitive list of AI terms (which have been graded according to their potential ambiguity) to tag the documents in the project abstracts in the funding database corpus. A key limitation of this process is that it is not possible to learn about the distinctive features of AI-related science published in non-AI journals unless the patterns are also present in AI journals.

The entire data corpus was cleaned[18] and tokenised (i.e. separated into 1 to 4-grams such as "robot", "deep learning", "natural language processing", and "knowledge representation and reasoning"). The tokens were vectorised to be compared in a vector space by mathematical measures of similarity, e.g. cosine similarity. This process generated a distributed representation of words or terms (also called "word embeddings" or continuous space representation of words). This has become a popular way of capturing distributional similarity (lexical, semantic, or even syntactic) between different words based on co-occurrence patterns. The basic idea is to

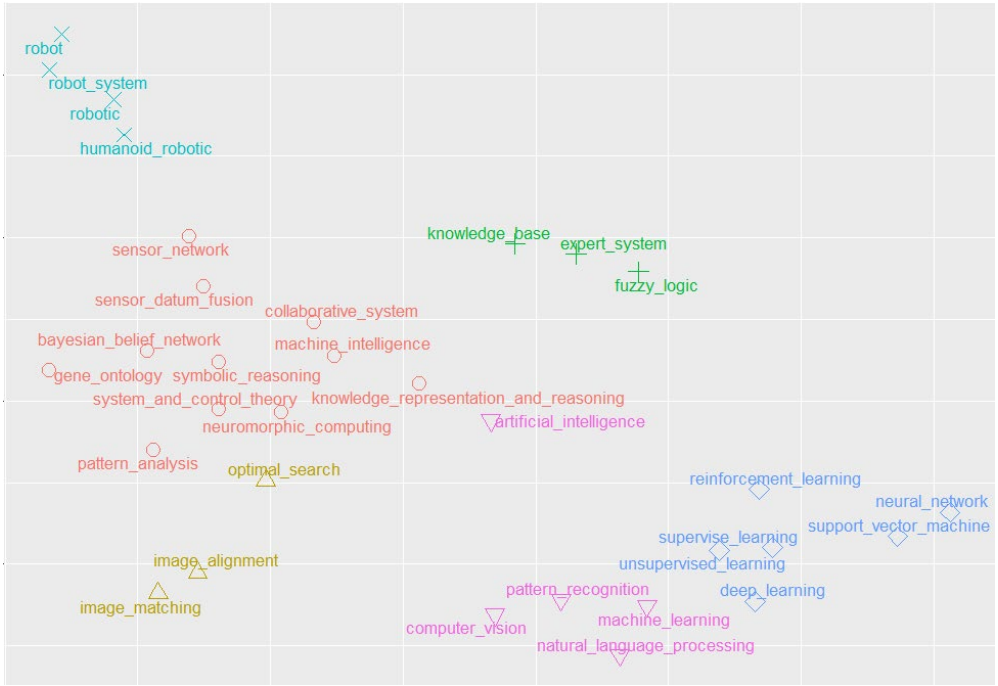
[18] The text was converted into lower case and lemmatised. Punctuation was removed, and all numbers were replaced by "0". Hyphenated terms were also processed.

represent each word in a vocabulary with a real-valued vector of some fixed dimension. This paper analysed the "corpus" of scientific publications in AI journals through a model of two-layer neural networks (Word2vec) trained to reconstruct linguistic contexts of words (Mikolov et al., 2013[24]). These networks "embedded" all of the terms in the corpus into 100-dimension vectors.

It is possible to examine the structure of the vector representations of the base 32 key term candidates through clustering analysis, as shown in **Figure 2.2**. This figure is a two-dimensional (2D) representation of the proximity of such terms in the 100-dimension vector space, with the terms clustered in six groups through a *k-means* algorithm. This visualisation of proximities across vector representations for these terms provides an indication of internal coherence. The cluster represented by plus signs "+" refers to AI methodologies used until the early 2000s but whose popularity has since waned. The "x" signs relate to robotics. Another cluster, represented by the diamond "◊", refers to types of automatic learning procedures. This cluster exhibits considerable proximity to the cluster (inverted triangles) including "machine learning" and common AI applications such as pattern recognition, natural language programming, and computer vision. The generic term "artificial intelligence" is likewise classified in this cluster and is positioned in a central spot in the 2D representation of terms. Image analysis terms form a distinctive cluster (upright triangles). The circle cluster, for its part, is more internally heterogeneous, containing in the same group statistical concepts used in AI, as well as terms relating to sensors and to gene ontology.

Figure 2.2. Cluster representation of base key AI terms from M- and C-lists

Two-dimensional cluster representation based on term embeddings in the Scopus AI-journal corpus

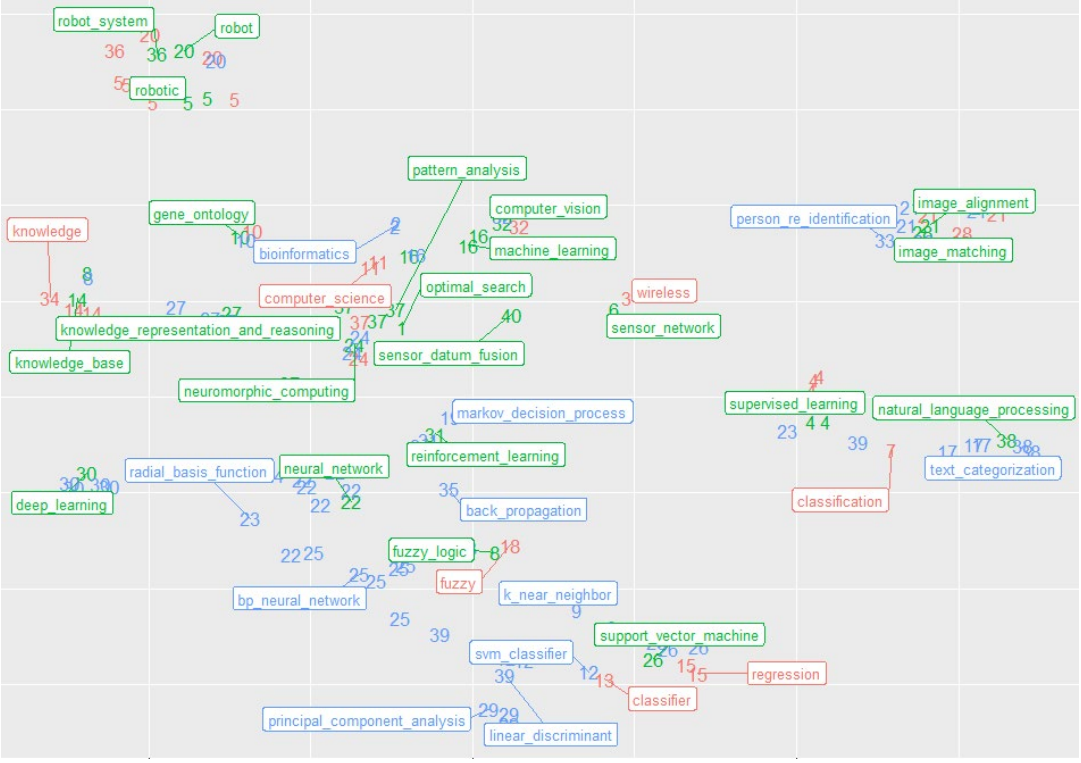


Note: The positions of the terms in the 2-dimensional space reflect the result of the clustering algorithm described in the main text and do not have a specific interpretation.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

Terms inside the AI-journal Scopus corpus with high[19] cosine similarity to the base terms found in the M-C list were selected as potential candidates to become additional key terms. The procedure yielded a total of 171 terms. 45 duplicated terms were removed or merged. For example, "machine learning techniques" (duplication with "machine learning") was removed. A total of 27 abbreviations were also removed to avoid duplication. The remaining 99 terms were subject to cluster analysis alongside the base key AI terms.

Figure 2.3. Cluster representation of base and additional AI terms extracted from the Scopus AI-journal corpus

Two-dimensional cluster representation based on "term embeddings" in the Scopus AI-journal corpus



Note: One single label per cluster is presented to facilitate readability. The colour coding is as follows: green for AI base terms, blue for terms selected as additional AI terms to be categorised into core or non-core, and red for removed terms that were not considered as AI-related in the subsequent scoring procedure.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018.

Visualising this larger set of terms in a two-dimensional space is more challenging but still possible through a more sparing use of labels, as shown in **Figure 2.3**.[20] This extended visualisation differs from the previous one as it represents connections across a broader set of terms, all derived from an entirely AI-related corpus (the AI journals). It helps identify the connections between base terms and co-occurring terms (that are

[19] The analysis set a minimum threshold of cosine similarity at +0.65. As the cosine similarity drops, the likelihood of retrieving terms that have irrelevant meanings rises considerably.

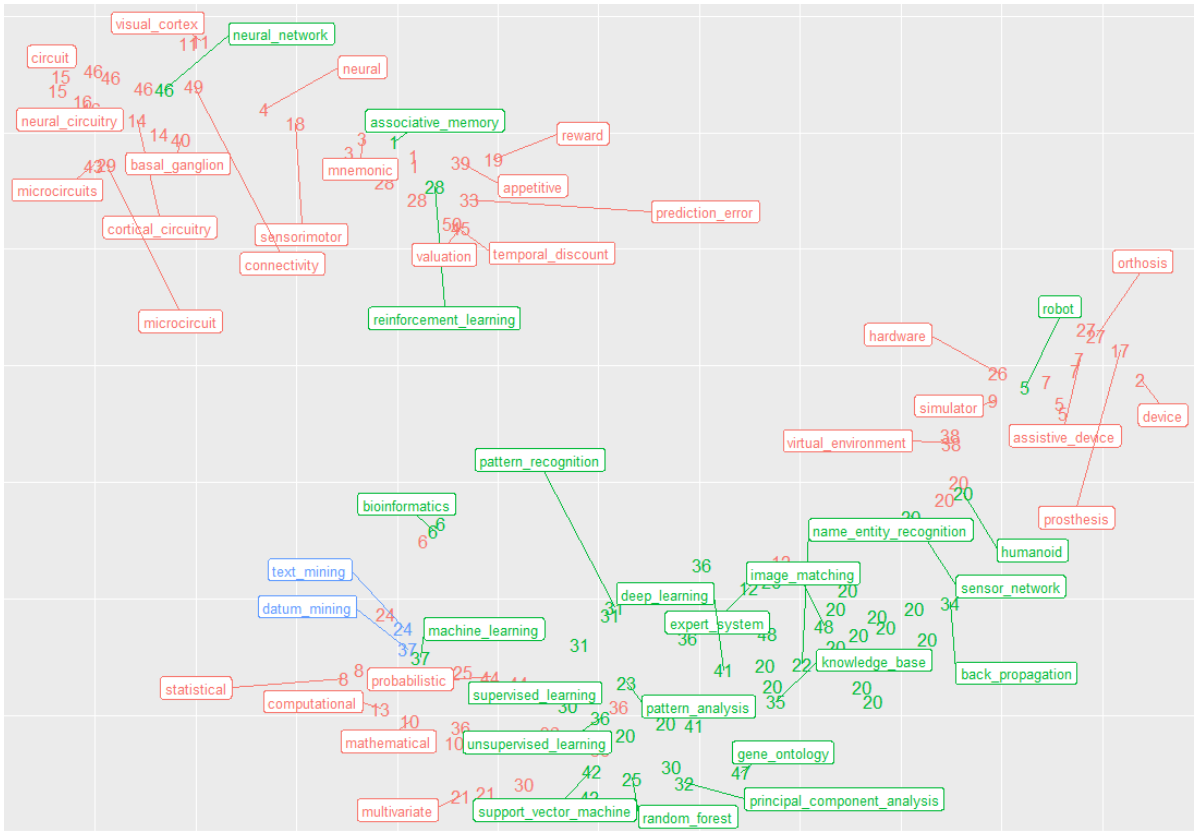
[20] More detail on the individual terms is available in the annex (**Table B.1)**.

likely to be AI terms themselves), as well as to those that may have an ambiguous connection to AI, particularly in corpora from other fields (e.g. wireless, classification, cognitive science, computer science, wheelchair, mosaicking, and registration). These latter terms were manually removed from the list to increase precision and to reduce the risk of false positives. This ultimately resulted in the base list of 32 terms being expanded by 65 additional terms.

Step 3: Further enrichment of key terms through the funding databases

Drawing on this extended list, the same methodology was applied to all 13 databases in order to assess the feasibility of identifying additional context-specific key AI terms, as well as to assess the interpretability of the key terms from an AI-relatedness perspective. This stage of the analysis resulted in only nine additional key terms. Overall, this resulted in 106 key AI terms, as presented in **Table B.2**, which also shows the corpus in which each term was first identified. The cases corresponding to the NIH (**Figure 2.4**) and to the NSF (**Figure 2.5**) are shown below.

Figure 2.4. Clustered vector representations of AI terms extracted from the NIH project funding corpus



Note: One single label per cluster is presented to facilitate readability, giving priority to the AI base term or in its absence the most frequent label in the cluster. The colour coding is as follows: green for AI base terms plus the terms retrieved from Scopus (extended list of terms), blue for terms identified as relevant additional AI terms, and red for terms that were not considered as AI-related and were excluded from the subsequent scoring procedure.
Source: OECD calculations based on NIH Reporter data, accessed December 2018.

In the case of NIH data, the visual representation shows two main groups of potentially AI-related terms[21]. The top left corner of **Figure 2.4** reveals the high level of ambiguity for AI tagging purposes of the "neural network" term, which is clustered around neuroscience-related terms. The same applies to two AI terms (reinforcement learning and associative memory), which appear to mostly be used in projects that investigate human cognitive processes. Although this might reveal the use of AI in (or connected to) the neuroscience domain[22], co-occurring terms in this space such as "neural", "prediction error", or "circuit" were nevertheless excluded from the list of key terms for document retrieval. Within the region that appears to be less ambiguously related to AI, one broad cluster is dominated by devices and hardware that tend to co-occur with references to the term "robot". A cluster of statistical terms is found to gravitate around the term "machine learning". These terms may also be weak signals of AI activity, but since there is a high risk that those terms are being used for standard epidemiological or biostatistics analysis, they were excluded from the list of AI terms.

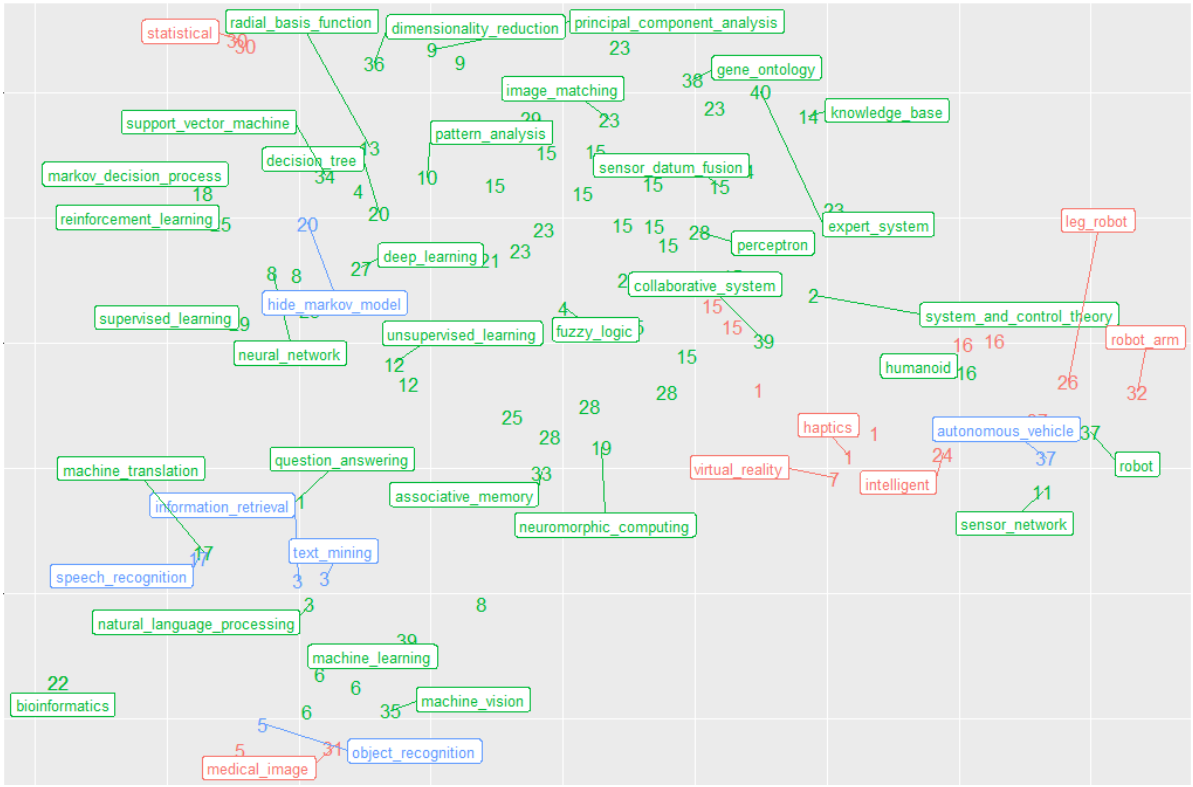
In the case of NSF data, the clustering exercise visualised in two dimensions in **Figure 2.5** presents a significantly larger set of AI terms given the database's coverage of core computer science fields. The clusters on display exhibit a varying degree of ambiguity with regards to their connection to AI. The bottom-left region of the figure appears to identify application areas for AI systems dealing with text, speech, and image recognition, which in turn link to medical and life sciences applications. The terms in the centre-right region relate to automation and robots, while the top-left region includes various types of machine learning techniques, which link to more generic statistical terms. The top-right corner incorporates AI strands unrelated to machine learning systems (expert systems, ontologies, and knowledge bases). The central space is more sparsely populated and consists of terms related to neuromorphic computing[23] and associative memory.

[21] Only the initial application document for a project was analysed in this step, as most of the subsequent application documents were duplicates of the initial one.

[22] In particular, research providing possible directions for artificially replicating how the human brain operates.

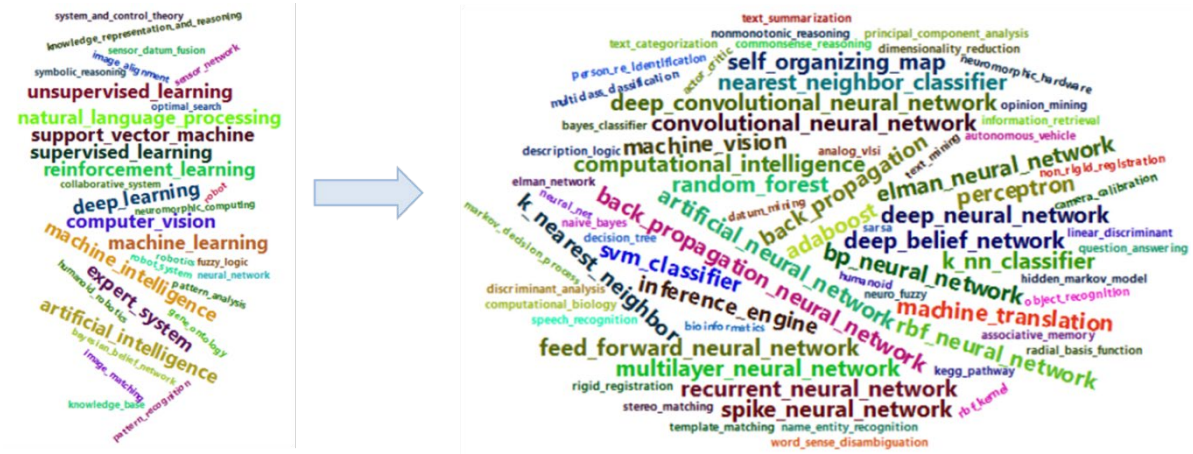
[23] An emerging interdisciplinary field focused on designing hardware/physical models of neural and sensory systems.

Figure 2.5. Clustered vector representations of AI terms extracted from the NSF project funding corpus



Note: One single label per cluster is presented to facilitate readability, giving priority to the AI base term or in its absence the most frequent label in the cluster. The colour coding is as follows: green for AI base terms plus terms from Scopus, blue for terms identified as relevant additional AI terms, and red for terms that are were not considered as AI-related and were excluded from the subsequent scoring procedure.
Source: OECD calculations based on NSF Award Search data, accessed December 2018.

Figure 2.6. Key term selection based on base list of AI terms



Base key AI terms from two key terms sets (MeSH and Cockburn)

Semi-automatically retrieved additional key AI terms from Scopus and funding databases

Note: This dual word cloud represents the extension from a base set of key AI terms based on seed "expert" lists to a text-mining extended list of AI terms for document retrieval. The full list of selected potential key AI terms is available in **Table B.2**. The key terms in the figure are lemmatised (e.g. machine learning -> machine learn). Core AI terms are presented using a larger font size.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018, and funding databases analysed.

Step 4: Labelling key terms as "core" and "non-core" to tag and classify AI-related projects

Equipped with a list of key terms, it was then possible to identify which documents incorporate such terms. The list enrichment process was oriented towards reducing the risk of a high false omission rate and the associated failure to identify relevant AI-related projects. As a result, further caution was needed to avoid selecting documents that utilize key AI terms but do not or are unlikely to involve the use or development of AI systems.

In addition to removing overly ambiguous terms from the candidate list, a very simple binary rating approach separated retained terms between "core" and "non-core" terms, according to the degree of ambiguity in their relationship to AI. The term "core" implies here "reliable" and refers to a low degree ambiguity in its relatedness with AI, rather than whether a term is foundational to methodology of AI (development vs applications). Furthermore, this paper's use of "core" has nothing to do either with the notion of "core AI" as AI platforms developed internally by organisations (i.e. coming from the core of the organisation) as opposed to the use of AI as a service.

This procedure is applied to terms in the base list as well as to the enriched terms. Some terms included in the expert M and C-lists such as "neural network" were categorised as non-core, because the term is not only used to describe an AI algorithm but also to describe a biological network of neurons. Similarly, "reinforcement learning" was classified as non-core as it can also be used in other R&D contexts.

2.3.2. Tagging documents with the list of key AI terms

The project selection criterion requires that a project description should contain (at minimum) either one core term or two or more distinct potentially ambiguous (non-core) terms.[24] In other words:

- A document was selected as (likely to be) AI-related if
  - At least one core key term was found within its title or abstract; or
  - Two or more distinct non-core terms were found (a special rule was applied to the terms "bioinformatics" and "computational biology"; in that case, an additional key term was required, as that particular term is likely to produce false positives)
- All other documents were classified as (likely to be) non AI-related

This is a rather simple and somewhat naïve procedure aimed at resolving potential ambiguity, as the implicit scoring and thresholds are defined somewhat arbitrarily. The idea is that by requiring the inclusion of at least two potentially ambiguous terms, the likelihood of accepting a non AI-relevant document will be significantly reduced. A higher threshold (or lower scoring for such terms) would certainly improve precision but would concurrently increase the risk of a high false omission rate by a substantial amount.

Another marked disadvantage of this naïve procedure is the extent of manual intervention required to assign terms into the three possible categories (AI core, AI non-core, and overly ambiguous). Ultimately, these challenges stem from the impossibility of training a selection algorithm due to the lack of a comprehensive, labelled database of projects.

In order to better understand the error resulting from the above procedure, a manual examination of potential biases was carried out by extracting and analysing four samples of documents:

- Within the set of documents identified as AI-related, 100 documents were randomly selected from each of the NIH and NSF corpora. These 200 documents were then examined to discover which documents might have been wrongly classified as AI-relevant (i.e. the approximate false discovery rate).
- Within the set of documents identified as not AI-related, 100 documents were selected from each of the NIH and NSF corpora. These 200 documents were then examined to discover which documents might have been wrongly classified as non AI-relevant (i.e. the approximate false omission rate).

The human examination of these 400 documents sought to establish whether the documents were unambiguously AI- or non AI-related, or whether it was impossible to tell either way without further information about each project. Within this latter class, it was often possible to distinguish between projects that laid out tasks for which AI tools were often required and others that did not provide an explicit connection to AI. This analysis provides an indication of the challenges associated with text mining

[24] To ensure distinctiveness, a number of terms were merged before the key term matching stage. For example, "humanoid robotics", "robot systems", and "robotic" were merged with the term "robot", and "neural net" was merged with "neural network".

project abstracts for information retrieval and classification according to AI-relatedness.

2.3.3. Topic modelling analysis of selected AI-related documents

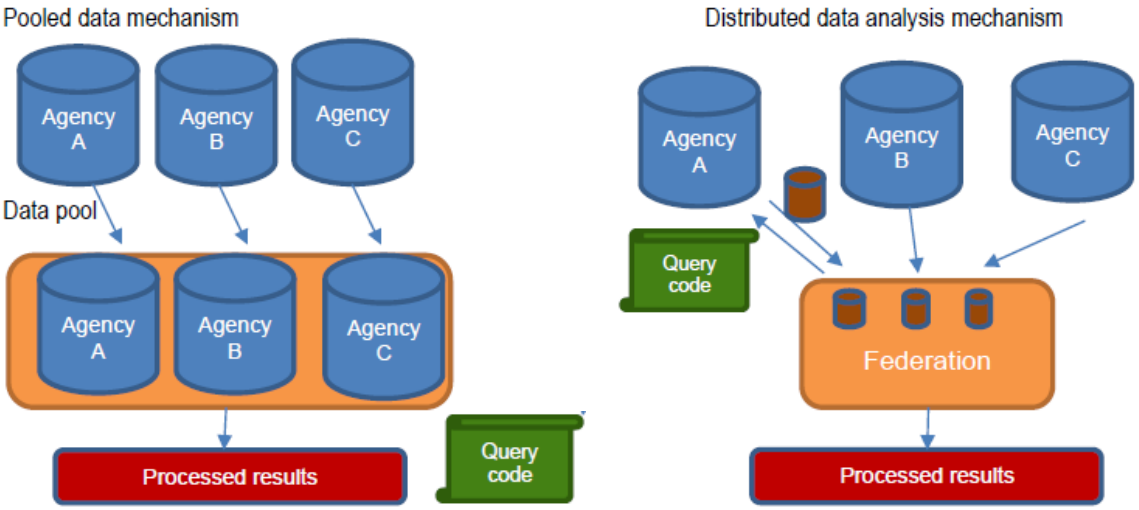
Based on the tagging of funding portfolios by AI-relevance, it is possible to examine what topics frequently appear in the texts (project titles and abstracts) associated with the tagged projects. Inference regarding what types of research are supported by the funding organizations (e.g. what technologies are often studied and for what purposes) is therefore possible.

Topic modelling based on Latent Dirichlet Allocation (LDA) was used for this analysis. LDA is a generative probabilistic model for collections of discrete data such as text corpora to find topics via a three-level hierarchical Bayesian model (Blei, Ng and Jordan, 2003[25]). To conduct LDA, the desired number of topics should be pre-defined – in this case, the number of topics were based on the coherence index, which represents the consistency of the topics, supplemented by manual checks of the multiple sets of topics (sets of 3, 6, 9, 12, 15, and 18 topics) generated by the algorithm for each database, as the index was not reliable enough to automatically decide the best number of topics.

2.3.4. Analysis under different data access regimes

As previously noted, not all funding databases are publicly available for download and analysis. This limitation prevented the implementation of the data pooling strategy described in **Figure 2.7** (left side). A distributed analysis mechanism was instead implemented, as illustrated in **Figure 2.7** (right side), through a collaborative engagement with the Dutch and Spanish delegations to NESTI, who were provided with the same Python code used for the databases directly pooled and analysed by the OECD secretariat. They implemented the code within their own systems and submitted the non-confidential statistical results back to OECD for inclusion in this report.

Figure 2.7. Illustration of centralised (pooled) and distributed data analysis mechanisms



Note: The distributed or federated mechanism does not pool all data but instead collects specific, previously agreed upon features from each agency-held database, using a query code that agencies run internally.
Source: OECD Fundstat project concept description paper.

The distributed analysis method enabled the inclusion of results for the Netherlands and Spain. In addition to this, the Spanish NESTI delegation undertook in parallel the analysis of the EC-funded projects available in CORDIS. This experience helps to demonstrate the feasibility of a broader and more inclusive federated/distributed model for Fundstat.

2.3.5. Analysing data in different languages

Another major step in the development of a multi-country infrastructure is addressing the language barriers. The databases under consideration are not all available in English. Four of them use different domestic languages, namely Dutch, French, Japanese, and Spanish. Such differences could cause problems, as the matching process required the languages of the key terms and the document to be the same. This analysis employed a manual translation approach for the key terms for Dutch, French, and Japanese, while a bulk machine translation of the project documents was applied in the case of the Spanish language documents in PlanEst. This choice was made by each team implementing the analysis directly on the data. Both approaches proved effective in delivering results for the purpose of this analysis but no comparison has yet been carried out which allows us to assess whether to recommend a given approach in the future.

3. Results

3.1. Key AI terms across funding databases

Figure 3.1 shows the lists of key AI terms that frequently appear in the successful R&D applications in each funding database. The occurrence of a key term in a particular database corresponds to the number of documents in the database in which that term appears at least once, normalised by the number of documents within each database for comparability. The top 10 most frequently occurring terms are listed for each database. "Robot" was the most frequently occurring term in many databases, with the exception of the databases associated with medical or life-science focused agencies. In these latter databases, "bioinformatics" appeared more frequently. Neither of these two terms are considered as core key AI terms, as they may also be used in non AI-related contexts.[25] Stronger key AI terms such as "machine learning" and "neural network" also appeared with high frequency in many databases.

Figure 3.1. Ten most frequent key AI terms, by funder/database

Number of documents in which term occurs, per 10 000 documents

Top 10 most frequently occurring key-terms	Funding organisations													Average
	AUS ARC	CAN CIHR	CAN NSERC	ESP PlanEst	FRA ANR	GBR GtR_Inno	GBR GtR_RC	JPN AMED	JPN KAKEN	NLD NWO	USA NIH	USA NSF	EU CORDIS	
robot	**84**	9	**53**	**178**	98	187	134	**166**	**88**	149	29	**257**	**196**	125
machine learning	41	15	50	42	71	**197**	**187**	61	37	**185**	34	239	122	99
bioinformatics	39	**35**	25	57	**162**	15	98	21	10	24	**201**	118	101	70
artificial intelligence	24	11	17	72	2	136	77	55	15	121	3	72	82	53
knowledge base	78	15	12	22	1	15	35	0	3	20	39	80	72	30
data mining	34	3	21	48	6	34	27	6	11	28	16	83	41	28
neural network	15	17	22	37	3	13	42	2	22	38	33	43	37	25
deep learning	7	1	8	5	26	32	40	10	19	96	3	38	22	24
computer vision	18	1	20	32	26	28	31	2	3	28	4	71	35	23
sensor network	22	0	25	39	2	26	28	0	7	11	0	86	35	22

Note: The most frequent term for each database has been bold highlighted underlined. NIH figures refer to individual granted applications. The NIH database contains multiple records for a number of projects that reflect separate funding decisions.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Figure 3.2 shows the occurrence of documents featuring selected key AI terms (machine learning, computer vision, natural language processing, deep learning, and

[25] Bioinformatics is considered in this context as weakly related to AI. Another frequent term, "Knowledge bases" is one of the MeSH AI terms and refers to "collections of facts, assumptions, beliefs, and heuristics that are used in combination with databases to achieve desired results, such as a diagnosis, an interpretation, or a solution to a problem". One example of a project in an AI context that refers to knowledge bases proposes "to develop two separate Decision Support Systems using totally different approaches, a heuristic model (knowledge based expert system) and a predictive statistical system. These Systems will be developed from a database of 3500-4000 MAG3 studies and will be designed to acquire the study, generate images and curves […], check for errors, extract the relevant quantitative data and then use these data to interpret the study." An example of a non AI-relevant project that refers to knowledge bases states that "training goals […] will enhance the applicant's knowledge base in child and adolescent mental health services".

expert system(s)). The term "machine learning" experienced a much faster growth in popularity than "expert system(s)" after 2010, indicating the growing dominance of the machine learning paradigm within AI R&D. The use of "expert system", initially a popular term, has flatlined or even declined. The accelerated growth of "deep learning" occurred relatively more recently, as is most readily noticeable in the NSF R&D funding figures. The figures for "computer vision" and "natural language processing", presented on a smaller scale, illustrate the growing importance of these two application areas of AI, with marked exceptions for computer vision in Australia (2014) and Spain (2008).

Figure 3.2. Occurrence of selected key AI terms in the 13 funding databases

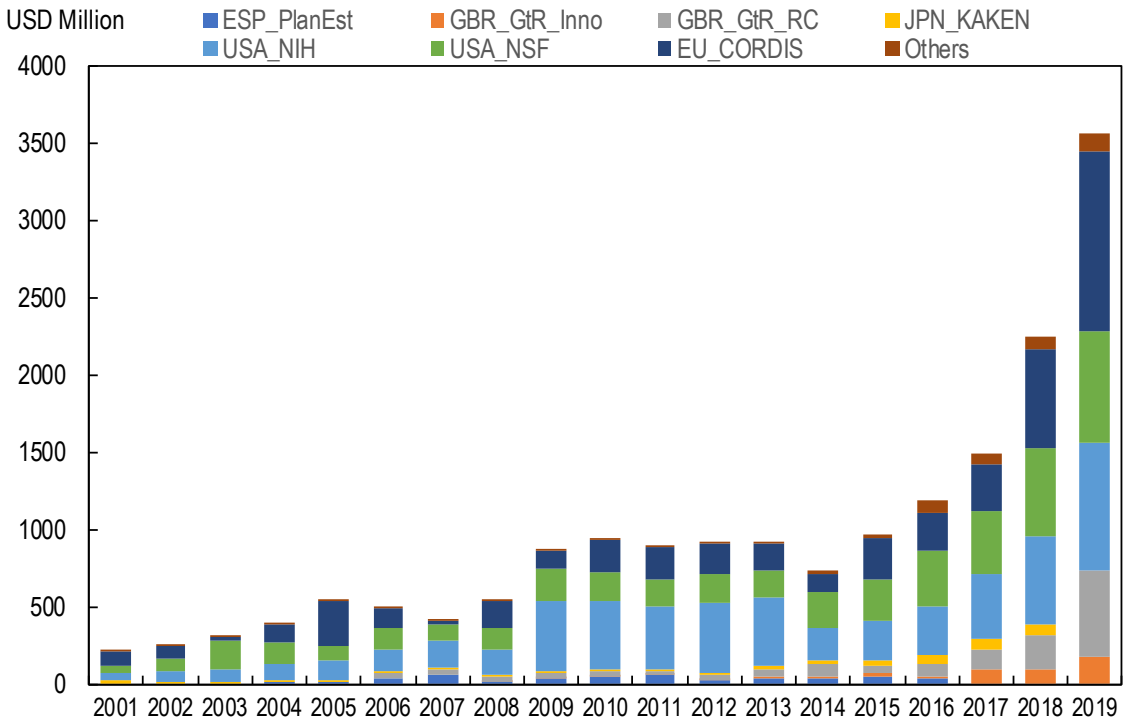
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

3.2. Estimates of AI-related R&D funding volumes

Estimates of AI-related R&D project funding for a targeted period are based on the selection procedure described in the previous section. As shown in **Figure 3.3**, the

analysis covers a total volume of AI-related R&D project funding that increased from USD 207 million in 2001 to nearly USD 3.6 billion in 2019 (figures in current prices, not adjusted for inflation and converted to USD through average exchange rates for the whole period). This estimate does not correspond to the actual growth, as information is not uniformly available for all years for a number of agencies and programmes. For the group of agencies with data available over a common and sufficiently long period (i.e. excluding Canada's NSERC, Spain's PlanEst, Japan's AMED, and the Netherland's NWO), the total volume of AI-related R&D project funding increased from USD 525 million in 2008 to USD 2 210 million in 2018.

Figure 3.3. Estimated AI-related R&D funding by selected agencies, 2001-19

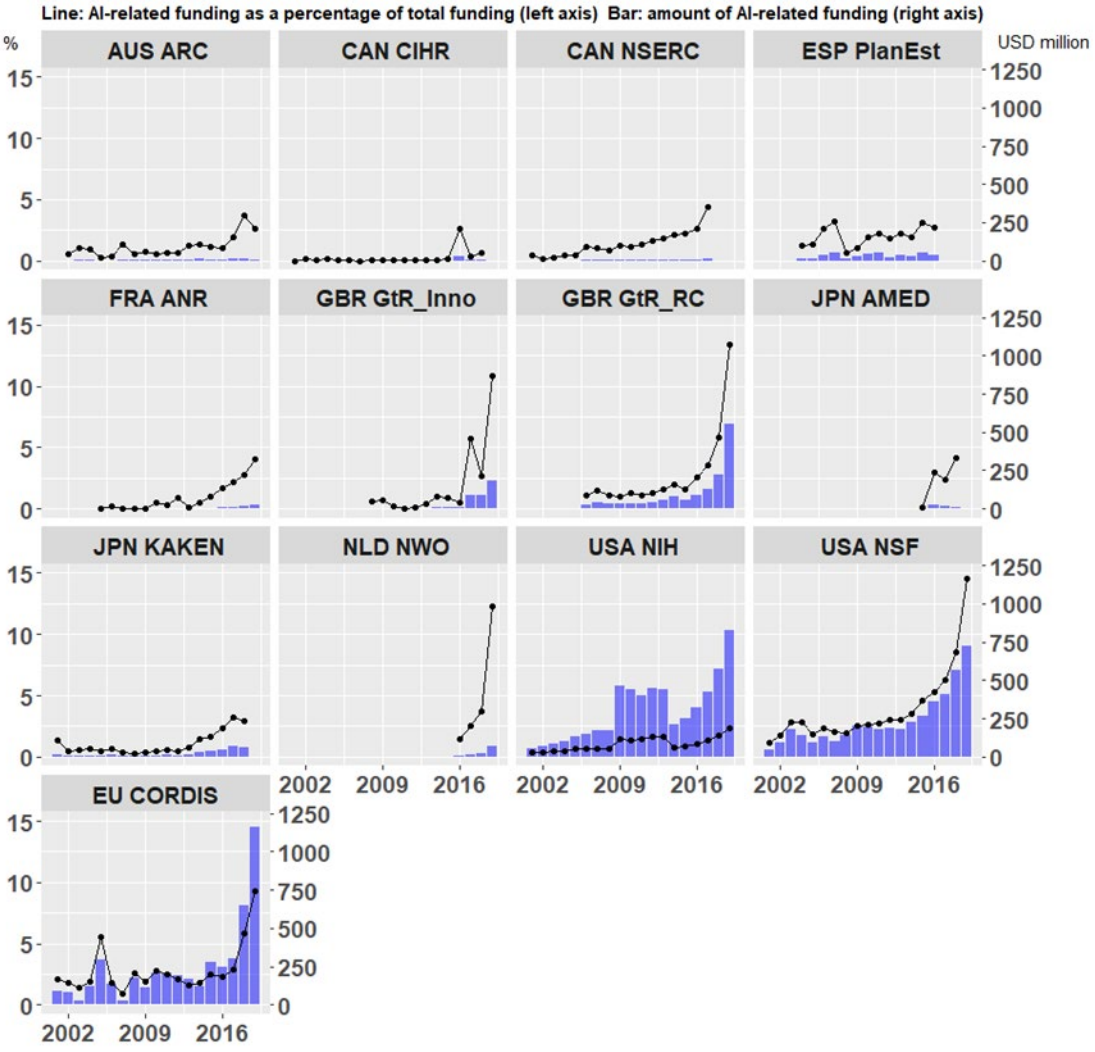


Note: The period of time for which data is available differs across funding agencies. ARC: 2002 to 2019, CIHR: 2001 to 2018, NSERC: 2001 to 2017, PlanEst: 2004 to 2016, ANR: 2005 to 2019, GtR_Inno (Innovate UK): 2008 to 2019, GtR_RC (Research councils): 2006 to 2019, AMED: 2015 to 2018, KAKEN: 2001 to 2018, NWO: 2016 to 2019, NIH and NSF: 2001 to 2019, CORDIS: 2001 to 2019.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Figure 3.4 shows overall trends in funding provided to AI-related projects by each funding agency. The bar charts (right axis) provide the agency specific funding volumes presented in **Figure 3.3**, while the line charts (left axis) illustrate the estimated proportion of AI-related funding out of total project funding over time. The USA NIH, the USA NSF, and the EU CORDIS are the largest AI R&D funders, followed by the UK Research Councils (GtR_RC). The GBR GtR Innovate UK (GtR_Inno), the GBR GtR_RC, the NLD NWO, and the USA NSF devote the highest proportions of their funding to AI R&D, each investing more than 10% of their total funding in recent years, with surges in funding occurring around 2013. AUS ARC,

CAN CIHR, CAN NSERC, ESP PlanEst, FRA ANR, JPN AMED and JPN KAKEN show relatively moderate upward funding trends.

Figure 3.4. Estimated AI-related R&D funding within selected agencies, 2001-2019



Notes: This is an experimental indicator.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

3.3. AI topics in R&D funded projects

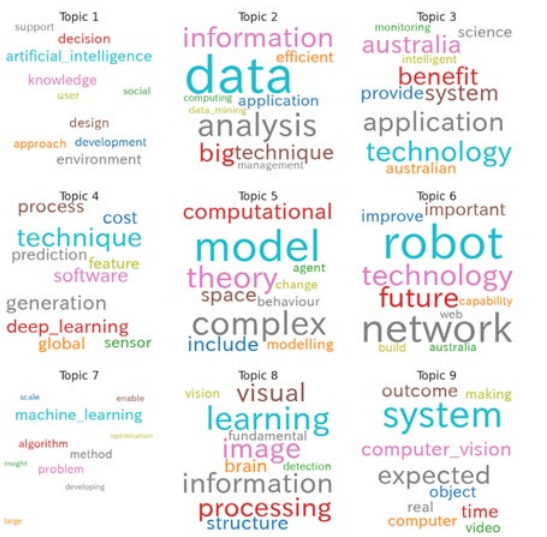
This section shows the results of the experimental topic modelling of AI-related documents selected using the key term matching approach. Topic modelling supplements the above analyses by undertaking a more fine-grained scrutiny of the topics that most frequently appear in the selected projects within each individual database.

The preliminary topic modelling results show useful information about the databases' key features (**Figure 3.5**). For example, a number of topics from the AI-related documents in AUS ARC included words associated with general AI techniques (Topic

4, 5, and 7 in **Figure 3.5(a)**). In the case of GBR GtR Innovate UK (**Figure 3.5 (b)**), words associated with applications of AI (e.g. business applications) appeared in several topics. The documents from the USA NIH database (**Figure 3.5(c)**) contained many topics related to medical care, while those from the USA NSF (**Figure 3.5 (d)**) covered a wide variety of topics. The results of topic modelling on other databases are shown in Annex C.

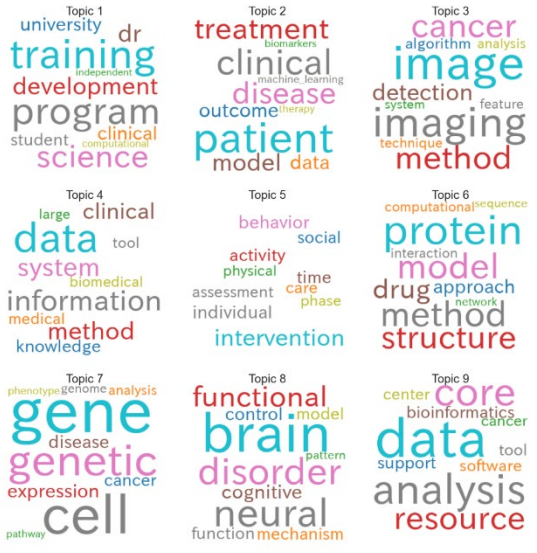
Figure 3.5. Topics from the AI-related documents by topic modelling analysis

(a) AUS ARC

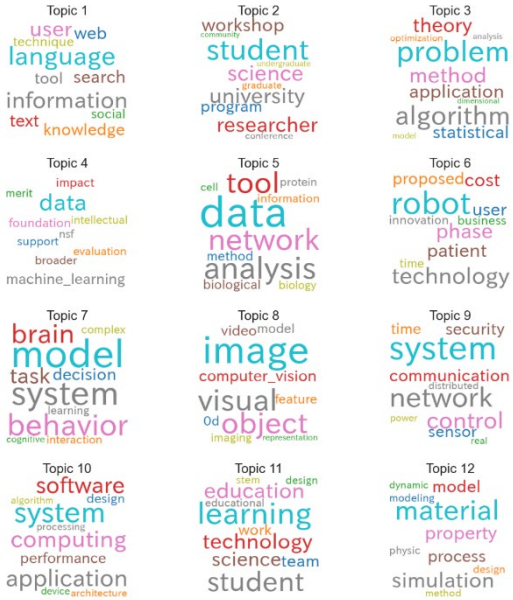


(b) GBR GtR Inno

(c) USA NIH



(d) USA NSF



Source: OECD calculations based on (a) ARC Grants Search data, (b) Gateway to Research (Innovate UK part), (c) NIH RePORTER data and (d) NSF Award Search data, accessed August 2020.

In order to compare the different funding streams, a manual labelling of topic subjects was undertaken for all 13 databases, based on the examination and interpretation of the terms present in each word cloud (**Table 3.1**). In this classification, five common themes and 21 common topics were selected. The five common themes were: *general AI techniques,* containing only itself as a topic; *AI prerequisites and impact,* which encompassed what is needed for AI to function but what is itself not strictly AI, and which contained the topics "education and training", "social impact", and "cost/production/monitoring", and "software development"; *AI fields*, containing the topics "computer vision/image or video processing", "NLP/text mining", "big data/data analysis", and "robots"; *AI application areas (non-medical),* containing the topics "business", "decision support", "network/service systems", "energy/power systems and devices", "smart technology", "social sciences", and "general/other applications"; and lastly *medical AI applications*, containing the topics "treatment and patients", "research", "diagnosis or imaging", "data", and "robots/devices".

Table 3.1. Classification of agency-specific topics into common themes and topics

The numbers within cell indicate the agency-specific topics assigned to common topics and themes

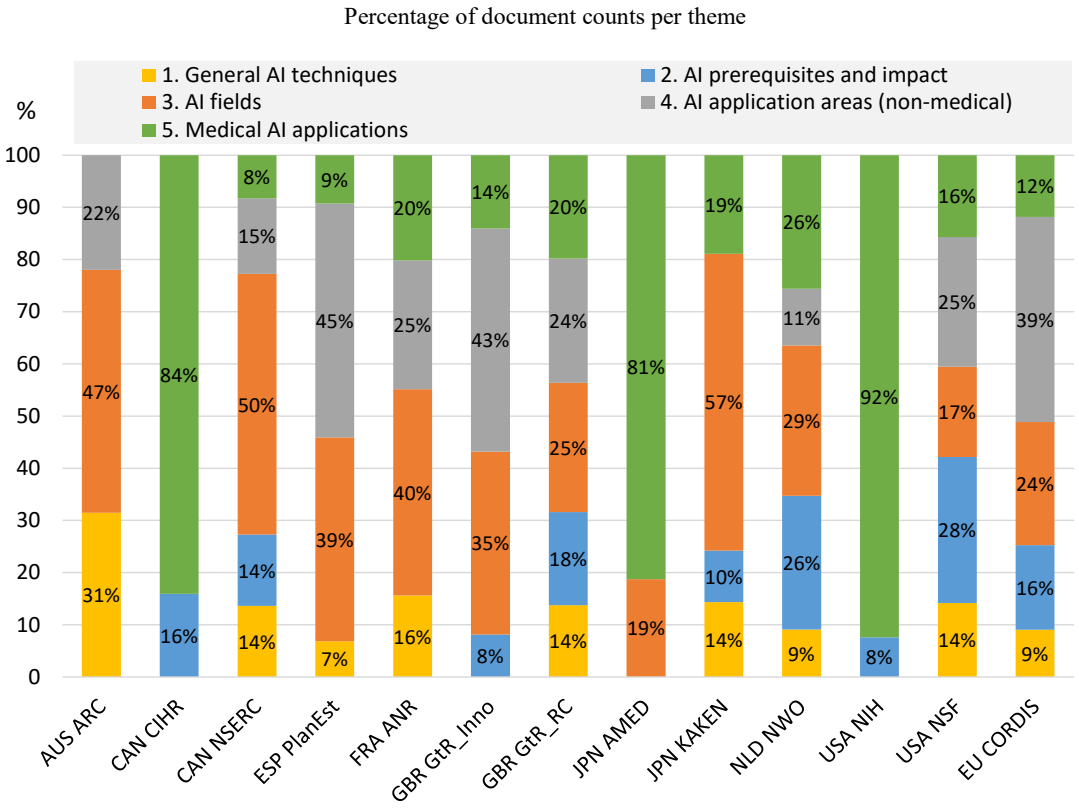
Common themes/topics	AUS ARC	CAN CIHR	CAN NSERC	ESP PlanEst	FRA ANR	GBR GtR_Inno	GBR GtR_RC	JPN AMED	JPN KAKEN	NLD NWO	USA NIH	USA NSF	EU CORDIS
1. General AI techniques													
1.1 General AI techniques	4; 5; 7		3; 10	1	9		8; 9		1	2		3	11
2. AI prerequisites and impact													
2.1 Education and training							11		3		1	2; 11	4
2.2 Social impact		1					12			5		4	
2.3 Cost/production/monitoring			11			1							1
2.4 Software development			12							7		10	
3. AI fields													
3.1 Computer vision/image or video processing	8; 9		8	2	1; 7; 8	5; 6	4	8	6			8	
3.2 NLP/text mining			7	7			3		9	1		1	3
3.3 Big data/data analysis	2		1; 9	5					5; 8	4			9
3.4 Robots	6		2	12	2	3	7		4	8			12
4. AI application areas (non-medical)													
4.1 Business				3		7							2
4.2 Decision support	1		4	8		9							
4.3 Network/service systems				6	4	4	1						5
4.4 Energy/Power systems and devices			6	4						3		9	7; 10
4.5 Smart technology	3			11		8							
4.6 Social sciences				9								7	8
4.7 General/other applications					6		2; 6					12	
5. Medical AI applications													
5.1 Treatment and patients (med)		2; 3; 4; 9	5	10	5	2	10	5; 9			2; 5		6
5.2 Research (med)		7					5	1; 2; 3; 7	2; 7	9	6; 7; 8		
5.3 Diagnosis or imaging (med)		8						4, 6		6	3		
5.4 Data (med)		5			3						4; 9	5	
5.5 Robots/devices (med)		6										6	
Number of topics by topic modelling analysis	9	9	12	12	9	9	12	9	9	9	9	12	12

Notes: Ad hoc topic classification based on preliminary topic modelling results. Word clouds corresponding to the topics in each database can be found in Annex C (**Figure C.2** (AUS ARC), **Figure C.4** (CAN CIHR), **Figure C.6** (CAN NSERC), **Figure C.8** (ESP PlanEst), **Figure C.10** (FRA ANR), **Figure C.12** (GBR GtR_Inno), **Figure C.14** (GBR GtR_RC), **Figure C.16** (JPN AMED), **Figure C.18** (JPN KAKEN), **Figure C.20** (NLD NWO), **Figure C.22** (USA NIH), **Figure C.24** (USA NSF), **Figure C.26** (EU CORDIS)). See Annex for details on each agency specific topic.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Figure 3.6 shows the distribution of document counts by the common themes within selected agencies. CAN CIHR, JPN AMED, and the USA NIH have a large share of documents that fall under the "medical AI applications" theme. AUS ARC, CAN NSERC, FRA ANR and JPN KAKEN have relatively high shares that fall under the

"AI fields" theme. Finally, more than 40% of all documents in ESP PlanEst and The GBR GtR Innovate UK (GtR_Inno) fall under the "AI application areas (non-medical)" theme.

Figure 3.6. Distribution of documents by common AI themes within selected agencies



Note: This is an experimental indicator. Topics under each theme are laid out in **Table 3.1**
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Table 3.2 shows the percentage of documents by the common topics within selected agencies. Under the "medical AI applications" theme, CAN CIHR has a large percentage of documents that share the common topic of "treatment and patients", while JPN AMED and the USA NIH have large percentages that fall under "research". In the case of the "AI fields" theme, the topics "computer vision/image or video processing" in AUS ARC, FRA ANR, and GBR GtR Innovate UK (GtR_Inno), and "big data/data analysis" in JPN KAKEN characterize over 20% of documents.

Table 3.2. Percentage of documents by common AI themes and topics within selected agencies

Percentage of document counts per common topic

Common themes/topics	AUS ARC	CAN CIHR	CAN NSERC	ESP PlanEst	FRA ANR	GBR GtR_Inno	GBR GtR_RC	JPN AMED	JPN KAKEN	NLD NWO	USA NIH	USA NSF	EU CORDIS
1. General AI techniques													
1.1 General AI techniques	31%	0%	14%	7%	16%	0%	14%	0%	14%	9%	0%	14%	9%
2. AI prerequisites and impact													
2.1 Education and training	0%	0%	0%	0%	0%	0%	9%	0%	10%	0%	8%	16%	8%
2.2 Social impact	0%	16%	0%	0%	0%	0%	9%	0%	0%	11%	0%	5%	0%
2.3 Cost/production/monitoring	0%	0%	7%	0%	0%	8%	0%	0%	0%	0%	0%	0%	8%
2.4 Software development	0%	0%	6%	0%	0%	0%	0%	0%	0%	14%	0%	8%	0%
3. AI fields													
3.1 Computer vision/image or video processing	23%	0%	13%	10%	26%	22%	9%	19%	12%	0%	0%	7%	0%
3.2 NLP/text mining	0%	0%	10%	12%	0%	0%	7%	0%	12%	7%	0%	10%	8%
3.3 Big data/data analysis	16%	0%	18%	9%	0%	0%	0%	0%	21%	10%	0%	0%	4%
3.4 Robots	8%	0%	9%	9%	13%	13%	9%	0%	12%	11%	0%	0%	11%
4. AI application areas (non-medical)													
4.1 Business	0%	0%	0%	12%	0%	10%	0%	0%	0%	0%	0%	0%	10%
4.2 Decision support	10%	0%	8%	6%	0%	11%	0%	0%	0%	0%	0%	0%	0%
4.3 Network/service systems	0%	0%	0%	8%	10%	8%	9%	0%	0%	0%	0%	0%	9%
4.4 Energy/Power systems and devices	0%	0%	6%	8%	0%	0%	0%	0%	0%	11%	0%	11%	13%
4.5 Smart technology	12%	0%	0%	3%	0%	14%	0%	0%	0%	0%	0%	0%	0%
4.6 Social sciences	0%	0%	0%	8%	0%	0%	0%	0%	0%	0%	0%	7%	7%
4.7 General/other applications	0%	0%	0%	0%	15%	0%	15%	0%	0%	0%	0%	7%	0%
5. Medical AI applications													
5.1 Treatment and patients (med)	0%	38%	8%	9%	12%	14%	11%	21%	0%	0%	20%	0%	12%
5.2 Research (med)	0%	15%	0%	0%	0%	0%	9%	44%	19%	13%	39%	0%	0%
5.3 Diagnosis or imaging (med)	0%	11%	0%	0%	0%	0%	0%	17%	0%	13%	11%	0%	0%
5.4 Data (med)	0%	10%	0%	0%	8%	0%	0%	0%	0%	0%	22%	7%	0%
5.5 Robots/devices (med)	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	9%	0%

Note: This is an experimental indicator.
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

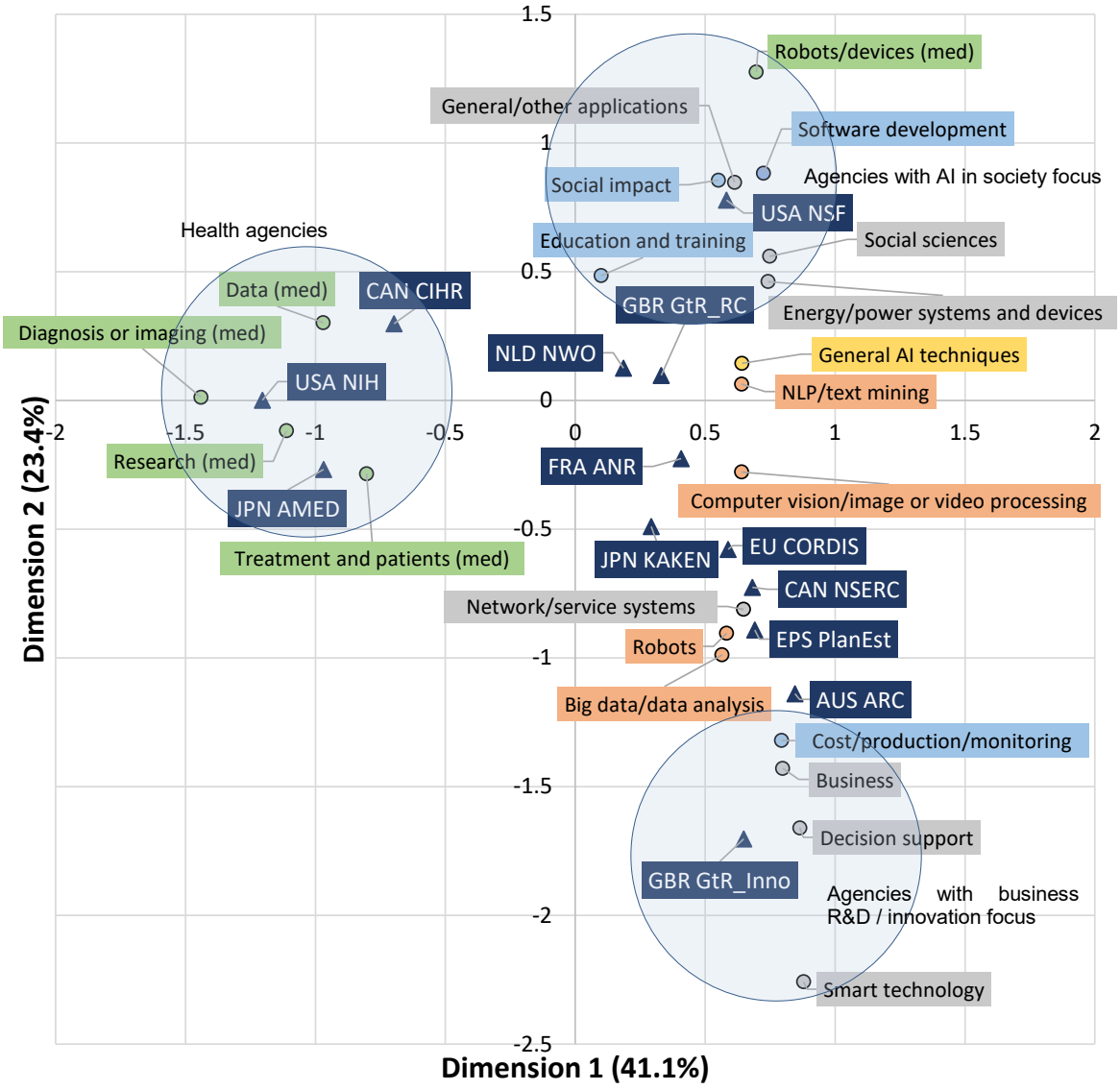
Correspondence analysis of the funding portfolios provides a more intuitive understanding of the similarities and characteristics of topics and agencies. **Figure 3.7** shows the visualised results of correspondence analysis in which the cross-tabulation table of AI-related document counts from 13 funding databases over 21 common topics were analysed and presented on a two-dimensional map. On this map, items with more distinctive features are placed far from the origin whereas those with greater commonality to the others are placed more centrally. In addition, items that are strongly related to each other with regards to their features are to be found in close proximity.

Analysing the position of the 21 common topics on the map, the horizontal axis (Dimension 1) can be interpreted as separating medically related databases and funding organisations (left quadrants) from their non-medical counterparts (right quadrants). The vertical dimension axis (Dimension 2) appears to distinguish between topics related to the "AI prerequisites and impact" theme, while many common topics related to "AI fields" or "AI application areas (non-medical)" themes are placed in the lower quadrants.

AI topic-wise, CAN CIHR, USA NIH, and JPN AMED are grouped on the left end of the horizontal axis, which reflects their status as medical agencies. It can also be construed that USA NSF funds projects related to "AI in society", while GBR GtR Innovate UK focuses on AI applications, especially those related to business R&D and innovation.

Figure 3.7. Correspondence analysis on the document counts per common topic

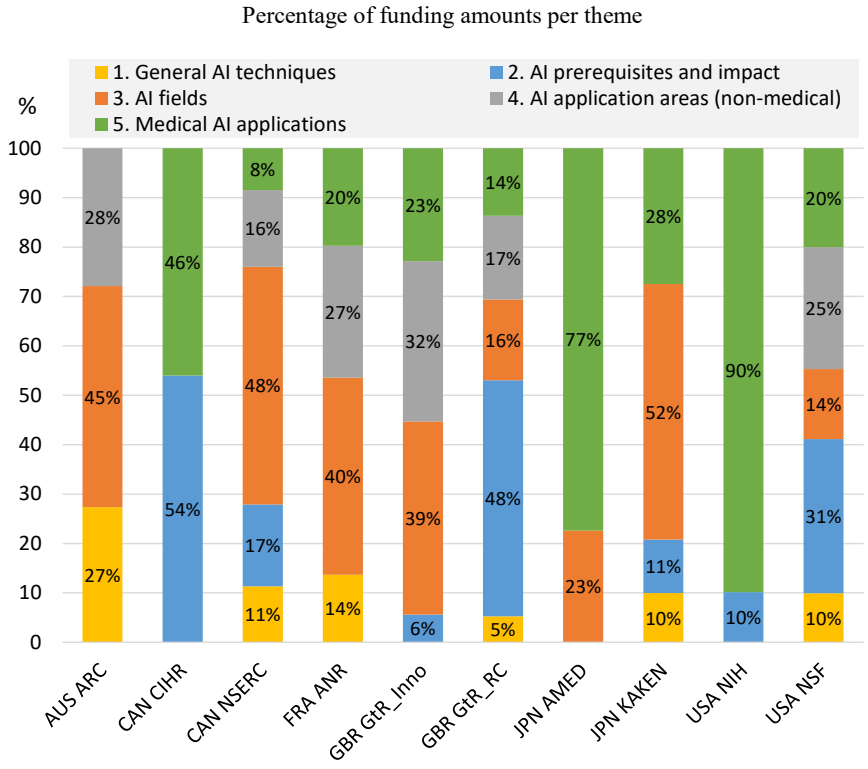
Two-dimension correspondence analysis plots of 13 funding databases and 21 common topics



Note: This correspondence analysis was carried out using the free software of "Real Statistics Using Excel (https://www.real-statistics.com/free-download/real-statistics-resource-pack/)". A two-dimension correspondence analysis (i.e. retaining just two factors) is sufficient provided the first two eigenvalues account for at least 50% of the variation, ideally much more. In this analysis, the value account for 64.5%, which corresponds to the sum of the contribution ratios in dimension 1 (41.1%) and 2 (23.4%).
Source: OECD, based on project microdata and results provided by the Netherlands and Spain.

Figure 3.8 shows the distribution of funding amounts by the common themes for the 10 agencies whose data was pooled by the OECD. The percentage of funding dedicated to the "AI prerequisites and impact theme" was higher than the equivalent percentage of documents for both CAN CIHR and the UK Research Councils (GtR_RC). There were no other major discrepancies between funding and count percentages.

Figure 3.8. Distribution of funding amounts by the common themes for 10 agencies



Note: This an experimental indicator. This result was analysed by only 10 funding databases whose data was pooled by the OECD.
Source: OECD, based on project microdata.

Table 3.3 shows the percentage of funding amounts by the common topics for 10 agencies. Under the "AI prerequisites and impact" theme, CAN CIHR accords a large percentage of its funding to projects that share the common topic of "social impact", while the UK Research Councils (GtR_RC) award a large percentage to research that falls under "education and training". In the case of CAN CIHR, it appears that one project, which received a particularly large amount of funding,[26] affected the results. The UK Research Councils (GtR_RC) gave substantial grants to several large projects[27] related to the "Education and training" common topic in 2019. These mega projects received around GBP 222 million (ca. USD 348 million) between them, which accounts for 24.4% of GtR RC's total AI-related funding over the whole period for which data is available (2006-2019).

[26] The project name is "Data Serving Canadians: Deep Learning and Optimization for the Knowledge Revolution" which has funding amounts of about CAD 28 million (ca. USD 23 million) and accounts for 51% of the total funding amounts of AI-related projects in 2001-2018, CAN CIHR.

[27] For example, the projects names are "EPSRC Hub in Quantum Computing and Simulation (https://gtr.ukri.org/projects?ref=EP/T001062/1)" and "Future Biomanufacturing Research Hub (https://gtr.ukri.org/projects?ref=EP/S01778X/1)", which have funding amounts of about GBP 24 million (ca. USD 37 million) and GBP 10 million (ca. USD 16 million), respectively.

Table 3.3. Percentage of funding amounts by the common topics for 10 agencies

Percentage of funding amounts per common topic

Common themes/topics	AUS ARC	CAN CIHR	CAN NSERC	FRA ANR	GBR GtR Inno	GBR GtR RC	JPN AMED	JPN KAKEN	USA NIH	USA NSF
1. General AI techniques										
1.1 General AI techniques	27%	0%	11%	14%	0%	5%	0%	10%	0%	10%
2. AI prerequisites and impact										
2.1 Education and training	0%	0%	0%	0%	0%	35%	0%	11%	10%	17%
2.2 Social impact	0%	54%	0%	0%	0%	13%	0%	0%	0%	5%
2.3 Cost/production/monitoring	0%	0%	8%	0%	6%	0%	0%	0%	0%	0%
2.4 Software development	0%	0%	9%	0%	0%	0%	0%	0%	0%	10%
3. AI fields										
3.1 Computer vision/image or video processing	25%	0%	12%	28%	12%	5%	23%	9%	0%	6%
3.2 NLP/text mining	0%	0%	9%	0%	0%	5%	0%	11%	0%	8%
3.3 Big data/data analysis	14%	0%	18%	0%	0%	0%	0%	18%	0%	0%
3.4 Robots	6%	0%	9%	12%	27%	7%	0%	14%	0%	0%
4. AI application areas (non-medical)										
4.1 Business	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%
4.2 Decision support	8%	0%	9%	0%	13%	0%	0%	0%	0%	0%
4.3 Network/service systems	0%	0%	0%	12%	5%	9%	0%	0%	0%	0%
4.4 Energy/power systems and devices	0%	0%	6%	0%	0%	0%	0%	0%	0%	10%
4.5 Smart technology	20%	0%	0%	0%	9%	0%	0%	0%	0%	0%
4.6 Social sciences	0%	0%	0%	0%	0%	0%	0%	0%	0%	6%
4.7 General/other applications	0%	0%	0%	15%	0%	7%	0%	0%	0%	8%
5. Medical AI applications										
5.1 Treatment and patients (med)	0%	13%	8%	11%	23%	8%	16%	0%	20%	0%
5.2 Research (med)	0%	12%	0%	0%	0%	6%	28%	28%	44%	0%
5.3 Diagnosis or imaging (med)	0%	8%	0%	0%	0%	0%	34%	0%	8%	0%
5.4 Data (med)	0%	6%	0%	9%	0%	0%	0%	0%	18%	11%
5.5 Robots/devices (med)	0%	7%	0%	0%	0%	0%	0%	0%	0%	9%

Note: This is an experimental indicator. This result was analysed by only 10 funding databases whose data was pooled by the OECD.
Source: OECD, based on project microdata.

4. Conclusions and next steps

This document has presented the procedures and main results of a text-based analysis of administrative data at the project-level on governmental R&D funding, a potentially very valuable source of information about the size and directionality of public investments in science and technology. These data, although limited in their usability due to access restrictions in some instances and limited scope and harmonisation, represent an invaluable tool for providing for timely insights on detailed aspects of government R&D funding.

The total volume of AI-related government R&D funding identified through this exercise grew from USD 207 million in 2001 to almost USD 3.6 billion in 2019, a seventeen-fold increase. While this might appear to represent a very large amount, and it does indeed represent a significant fraction of the funding streams that have analysed, this sum may be dwarfed by the sheer level of business R&D investment that appears to be taking place in parallel. This is the case if one considers that a single heavily AI-reliant company like Alphabet reported USD 20 billion worth of total R&D expenses in 2018, and the United States National Center for Science and Engineering Statistics has recently provided a conservative estimate of business AI R&D investment in the order of USD 9 billion out of its official estimate of over USD 160 billion worth of R&D on software products and software embedding technologies (NCSES, 2020[26]).

While countries develop their own mechanisms for providing official statistics on AI R&D expenditure and funding, this quantitative case study, in which a series of relatively simple text mining methods were applied to funding data from 13 funding databases in eight countries and a region, highlights the potential for using project level data descriptions to carry out in-depth, internationally coordinated analysis of the content and methods of R&D with such type of funding data. The findings on the subject of interest, AI-related R&D, confirm widely reported upward trends in R&D funding and provide additional insights about the topics of AI R&D that funding agencies around the world are investing in. This analysis could in the future be extended to other agencies and countries, and monitor developments over time. These lessons can also be particularly helpful in guidance future survey-based efforts, including those targeted at R&D being carried within firms.

This work has also represented a pilot exercise in assessing the feasibility of constructing a multi-country infrastructure on R&D project funding for analytical purposes ("*Fundstat*"), which would include the identification of emerging R&D domains and application areas, in light of heightened interest in the directionality of R&D support by governments.

This data-driven "classification and measurement" case study identifies challenges such as finding the right balance between mechanical procedures and individual judgements at different stages of the data management and analysis process. Data-based classification decisions regarding AI or other forms of emerging and enabling technologies also require knowledge of when applications of such technology become common enough to longer warrant detailed descriptions in project abstracts. Moreover, to be operationalised in a given corpus, existing definitions of AI will necessarily be context dependent.

The study reveals that funding microdata can shed multiple insights on the structure of R&D spending by governments. Crucially, information on funding gives a true measure of the scale of a project as opposed to simple document counts, while project information can have a very significant lead time over research output data. Data integrity of project descriptions is essential. This type of analysis is reliant on a comprehensive data infrastructure with meaningful and informative project descriptions. Text-mining of data containing superficial descriptions will fail to identify key features of projects, especially descriptions of project methodologies, which can be critical for topic relevance identification purposes. Ideally, it would better to conduct the analysis on the full body of project descriptions but access to these may not be possible. It should also be borne in mind that if data-driven insights are used to incentivise application and granting behaviour, the informational content in project descriptions will likely change as a result and so will be the conclusions that can be drawn from the analysis.

This work has in particular demonstrated that it is not strictly necessary --although it might be desirable in some instances where transparency overrides confidentiality considerations-- for all project-level data to be in the public domain. Decentralised, collaborative distributed approaches for the analysis of project level data have proved to be feasible as demonstrated through Dutch and Spanish funding data. However, further advances in data harmonisation are likely to be necessary in order to enable future longitudinal and cross-country analysis. Different and evolving patterns of project descriptions can lead to marked differences in results of text mining approaches. Investigating these issues will be part of the remit of the newly established

OECD Expert Group on the Measurement and Analysis of R&D and innovation administrative data (MARIAD) to assist NESTI in the pursuit of and quality assessment of this type of statistical analysis.

AI is far from being the only research field that evades easy definitions but whose emergence remains critical to track. This pilot study will serve as a prototype for this new OECD expert group to take on the development of broader analysis mechanisms capable of assessing government contributions to a myriad of fields and applications, including pandemic resilience objectives and outcomes connected with the UN Sustainable Development Goals.

Last but not least, it is worth noting the potential contributions of this line of work in informing and assisting survey-based measurements. For instance, mixed methods can be deployed that combine data-driven solutions with surveys. The two approaches complement each other: for example, surveys can allow for the construction of better text mining algorithms. Compared to data-driven approaches, surveys can also be more easily modified so as to best target actors engaged in more narrowly defined fields.

References

Abadi, H., Z. He and M. Pecht (2020), "Artificial Intelligence-Related Research Funding by the U.S. National Science Foundation and the National Natural Science Foundation of China", *IEEE Access*, Vol. 8, pp. 183448-183459, http://dx.doi.org/10.1109/access.2020.3029231. [19]

Acharya, A. and Z. Arnold (2019), *Chinese Public AI R&D Spending: Provisional Findings*, Center for Security and Emerging Technology, https://cset.georgetown.edu/research/chinese-public-ai-rd-spending-provisional-findings/ (accessed on 6 January 2021). [5]

Annapureddy, A. et al. (2020), "The National Institutes of Health funding for clinical research applying machine learning techniques in 2017", *npj Digital Medicine*, Vol. 3/1, http://dx.doi.org/10.1038/s41746-020-0223-9. [20]

Baruffaldi, S. et al. (2020), "Identifying and measuring developments in artificial intelligence: Making the impossible possible", *OECD Science, Technology and Industry Working Papers*, No. 2020/05, OECD Publishing, Paris, https://dx.doi.org/10.1787/5f65ff7e-en. [18]

BEIS and DCMS (2018), *Industrial Strategy Artificial Intelligence Sector Deal*, https://www.gov.uk/government/publications/artificial-intelligence-sector-deal. [7]

Blei, D., A. Ng and M. Jordan (2003), *Latent Dirichlet Allocation*, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed on 12 October 2020). [25]

Bode, C. et al. (2019), "A Guide to the Dimensions Data Approach A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it Contents", http://dx.doi.org/10.6084/m9.figshare.5783094. [15]

Cockburn, I. et al. (2018), *The Impact of Artificial Intelligence on Innovation*, http://www.nber.org/papers/w24449. [17]

EC (2018), *Artificial Intelligence for Europe; 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; COM(2018) 237 final*, https://knowledge4policy.ec.europa.eu/publication/communication-artificial-intelligence-europe_en. [6]

Freyman, C., J. Byrnes and J. Alexander (2016), "Machine-learning-based classification of research grant award records", *Research Evaluation*, Vol. 25/4, pp. 442-450, http://dx.doi.org/10.1093/reseval/rvw016. [29]

Gallo, F. et al. (2020), *Biomedical and health research: an analysis of country participation and research fields in the EU's Horizon 2020*, Springer Science and Business Media B.V., http://dx.doi.org/10.1007/s10654-020-00690-9. [32]

Goldstein, N. (1992), "The defense advanced research projects agency's role in artificial intelligence R&D: Case study of the military as the national agent for technological and industrial change", *Defense Analysis*, Vol. 8/1, pp. 61-80, http://dx.doi.org/10.1080/07430179208405524. [3]

Hajkowicz SA et al. (2019), *ARTIFICIAL INTELLIGENCE Solving problems, growing the economy and improving our quality of life*, CSIRO Data61, https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap (accessed on 17 September 2020). [10]

Kawamura, T. et al. (2018), "Funding map using paragraph embedding based on semantic diversity", *Scientometrics*, Vol. 116/2, pp. 941-958, http://dx.doi.org/10.1007/s11192-018-2783-x. [27]

MESRI (2020), *L'état de l'Enseignement supérieur, de la Recherche et de l'Innovation en France n° 13*, https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/EESR-FR.pdf (accessed on 11 January 2021). [22]

MESRI and DINUM (2018), *Stratégie nationale de recherche en IA*, https://cache.media.enseignementsup-recherche.gouv.fr/file/strategie_IA/60/7/mesri_IA_dep_A4_09_1040607.pdf (accessed on 12 October 2020). [8]

MICINN (2019), *SPANISH RDI STRATEGY IN ARTIFICIAL INTELLIGENCE*, Ministry of Science, Innovation and Universities, https://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial_EN.PDF (accessed on 18 September 2020). [9]

Mikolov, T. et al. (2013), "Distributed Representations of Words and Phrases and their Compositionality", *arXiv:1310.4546*, https://arxiv.org/abs/1310.4546v1. [24]

National Science Foundation (2020), *Survey of Federal Funds for Research and Development Fiscal Years 2018–19*, https://ncsesdata.nsf.gov/fedfunds/2018/ (accessed on 23 November 2020). [21]

NCSES (2020), *Business Research and Development: 2018*, National Center for Science and Engineering Statistics. NSF 21-312., Alexandria, VA: National Science Foundation, https://ncses.nsf.gov/pubs/nsf21312/ (accessed on 17 June 2021). [26]

NRC (1999), *Funding a Revolution: Government Support for Computing Research*, The National Academies Press, Washington, DC, http://dx.doi.org/10.17226/6323. [2]

NSTC (2020), "FY2021-NITRD-Supplement", https://www.nitrd.gov/pubs/FY2021-NITRD-Supplement.pdf (accessed on 18 September 2020). [4]

OECD (2021), *OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity*, OECD Publishing, Paris, https://dx.doi.org/10.1787/75f79015-en. [14]

OECD (2020), "The effects of R&D tax incentives and their role in the innovation policy mix: Findings from the OECD microBeRD project, 2016-19", *OECD Science, Technology and Industry Policy Papers*, No. 92, OECD Publishing, Paris, https://dx.doi.org/10.1787/65234003-en. [16]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, OECD-LEGAL-0449, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [1]

OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, https://doi.org/10.1787/d62f618a-en. [23]

OECD (2018), "Blue Sky perspectives towards the next generation of data and indicators on science and innovation", in *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-19-en. [13]

OECD (2015), "Concepts and definitions for identifying R&D", in *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264239012-4-en. [12]

Park, J. et al. (2016), *Analyzing NIH Funding Patterns over Time with Statistical Text Analysis*, the Association for the Advancement of Artificial Intelligence, Phoenix, Arizona USA, https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/download/12648/12442 (accessed on 16 September 2020). [33]

Perrault, R. et al. (2019), *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf (accessed on 6 January 2021). [11]

Scarpelli, M., B. Whelan and K. Farahani (2020), "Domain Classification and Analysis of National Institutes of Health Funded Medical Physics Research", *Medical Physics*, http://dx.doi.org/10.1002/mp.14469. [31]

Schmutz, A. et al. (2019), "Mapping the global cancer research funding landscape", *JNCI Cancer Spectrum*, Vol. 3/4, http://dx.doi.org/10.1093/jncics/pkz069. [30]

Talley, E. et al. (2011), *Database of NIH grants using machine-learned categories and graphical clustering*, http://dx.doi.org/10.1038/nmeth.1619. [28]

Annex A. Overview of R&D funders and databases

ARC (AUS)

The **ARC** (Australian Research Council) is an Australian funding organisation providing funding to basic and applied research and training projects across all disciplines, although clinical and other medical research is mainly supported by a separate funding organisation, the National Health and Medical Research Council. The ARC provides a project database named ARC Grants Search, in which data dating from 2002 to 2019 (26 677 projects, which received a total of USD 8 994 million) was available when accessed in August 2020.

CIHR (CAN)

CIHR (the Canadian Institutes of Health Research) is a Canadian federal funding organisation dedicated to health-related R&D projects. CIHR provides a project database, the Canadian Research Information System, whose data is downloadable via the open data portal of the Canadian government. The data downloaded in August 2020 consisted of 56 778 projects, which dated from between 2001 and 2018 and which received USD 14 147 million.

NSERC (CAN)

NSERC (the Natural Sciences and Engineering Research Council) is a Canadian federal funding organisation dedicated to R&D projects in natural sciences and engineering. NSERC provides a project database called NSERC's Awards Database, whose data is downloadable via the open data portal of the Canadian government. The data downloaded in August 2020 consisted of 175 945 projects from 2001 to 2017, which received a total of USD 3 402 million.

PlanEst (ESP)

The database of the **PlanEst** (National Plan for Scientific and Technological Research and Innovation) is an administrative funding database of the Spanish Ministry of Science and Innovation. It contains project data from four funding organisations[28] for basic research, applied research, and technological development projects. The Spanish Foundation for Science and Technology (FECYT) analysed the data accessed in July 2017 (they were collected by the Ministry at different time points) with the assistance of Jerónimo Arenas (University Carlos III of Madrid). The data consisted of 67 770 projects, which received USD 22 256 million between 2004 and 2016.

ANR (FRA)

The **ANR** (Agence Nationale de la Recherche) is a French funding organisation founded in 2005, providing funding to R&D projects on basic and targeted research, technological innovation, technology transfer, and public-private partnerships. The

[28] The four funding organisations are: State Research Agency, Centre for the Development of Industrial Technology, Institute of Health Carlos III, and Secretariat of State for Digitisation and Artificial Intelligence.

ANR provides a project database called Appels à projets ANR, whose data is downloadable via the open data portal of the French government. The data downloaded in August 2020 consisted of 20 123 projects, which received USD 6 506 million between 2005 and 2019.

GtR (GBR)

The **GtR** (Gateway to Research) is a UK funding database containing R&D projects funded by Innovate UK, by the National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs), and by seven additional research councils. [29] Unlike the eight other funding organisations, which focus on R&D, Innovate UK funds innovation projects; for this reason, our analysis separated Innovate UK data from all other data (dubbed Research Councils data). Since 2018, Research Councils UK and Innovate UK are part of the newly created agency Research and Innovation UK.

The data is downloadable from the GtR website. The data downloaded in August 2020 consisted of 18 424 projects (Innovate UK data), which received USD 14 281 million between 2008 to 2019, and 80 736 projects (Research Councils data), which received USD 46 280 million between 2006 to 2019.

AMED (JPN)

The **AMED** (Japan Agency for Medical Research and Development) was founded in 2015 and is dedicated to funding to health research. The AMED provides a project database called AMEDfind, which does not have download function. The organisation kindly offered a bulk data version of the system in March 2020, which contained 4 765 projects, totalling USD 4 213 million between 2015 and 2018.

KAKEN (JPN)

The JSPS (Japan Society for the Promotion of Science) funds scientific research in a wide range of fields. The JSPS provides a project database called **KAKEN** (Database of Grants-in-Aid for Scientific Research), through which their project data are downloadable. The data downloaded in August 2020 consisted of 466 709 projects, which received USD 33 750 million between 2001 to 2018.

NWO (NLD)

The **NWO** (Dutch Research Council) funds a wide range of scientific research and applied and engineering sciences projects. The NWO holds a project database, which is not open to the public. The data was analysed by the NWO using the code developed by the OECD. The Rathenau Institute analysed data accessed in March 2020. The data covered 7 177 projects, which received USD 2 186 million between 2016 to 2019.

[29] These seven research councils are: 1. Arts and Humanities Research Council (AHRC); 2. Biotechnology and Biological Sciences Research Council (BBSRC); 3. Economic and Social Research Council (ESRC); 4. Engineering and physical Sciences Research Council (EPSRC); 5. Medical Research Council (MRC); 6. Natural Environment Research Council (NERC); and 7. Science and Technology Facilities Council (STFC).

NIH (USA)

NIH (National Institutes of Health) is the largest biomedical research funding organisation in the world. Human health is the primary objective of all research NIH funds. NIH is made up of 27 institutes and centres (ICs), 24 of which can provide grant awards. These ICs award more than 80% of the NIH budget each year to support investigators across universities, medical schools, and other research organisations around the world. Around 10% of NIH's budget supports internal research and scientific activity.

NIH provides a database called NIH RePORTER, whose data is downloadable from NIH's ExPORTER website. The data downloaded in August 2020 consisted of 1 428 472 projects, which received USD 497 955 million between 2001 to 2019.

NSF (USA)

NSF (National Science Foundation) is the principal US federal agency in charge of supporting civil R&D across all fields of fundamental science and engineering, with the exception of the medical sciences, which is the domain of NIH.[30] Unlike NIH, NSF does not have its own internal R&D activity. As the organisation responsible for funding research on engineering and computer and information science, NSF supports several projects pushing the boundaries of AI, in addition to supporting projects across several disciplines that may make varying use of AI in their work.

The NSF provides a database called NSF Award Search, whose data is downloadable from its website. The data downloaded in August 2020 consisted of 224 307 projects, which received USD 114 883 million between 2001 to 2019.

CORDIS (EU)

CORDIS (Community Research and Development Information Service) is an administrative funding database belonging to the European Commission. It contains project data for the EU's framework programmes like Horizon 2020 and FP7. The Spanish Foundation for Science and Technology (FECYT) analysed the data accessed in April 2020 with the assistance of Jerónimo Arenas (University Carlos III of Madrid). The data consisted of 72 061 projects, which received USD 142 864 million between 2001 to 2019.

[30] Other major US R&D funding organisations with R&D funding levels higher than NSF are the Department of Defense (DOD, the largest by far in terms of funding), NASA, and the Department of Energy (DOE).

Annex B. Key terms selection

Table B.1. Clustering and treatment of AI-related terms in the AI-journal corpus

Cluster number	Quasi-synonyms or key AI terms	Status
1	OPTIMAL_SEARCH	Key AI term
2	bioinformatics	Selected as key AI term
2	computational_biology	Selected as key AI term
3	wireless	Removed
4	semi_supervised	Removed
4	supervised	Removed
4	transductive	Removed
4	unsupervised	Removed
4	SUPERVISED_LEARNING	Key AI term
4	UNSUPERVISED_LEARNING	Key AI term
5	autonomous	Removed
5	drone	Removed
5	mechatronic	Removed
5	rover	Removed
5	teleoperated	Removed
5	HUMANOID_ROBOTIC	Key AI term
5	ROBOTIC	Key AI term
6	ad_hoc_network	Removed
6	SENSOR_NETWORK	Key AI term
7	classification	Removed
8	inference_engine	Selected as Key AI term
8	EXPERT_SYSTEM	Key AI term
8	FUZZY_LOGIC	Key AI term
9	k_near_neighbor	Selected as key AI term
9	naive_bayes	Selected as key AI term
10	kegg_pathway	Selected as key AI term
10	protein_protein_interaction	Removed
10	GENE_ONTOLOGY	key AI term
11	cognitive_science	Removed
11	computer_science	Removed
12	bayes_classifier	Selected as key AI term
12	k_nn_classifier	Selected as key AI term
12	near_neighbor_classifier	Selected as key AI term
12	svm_classifier	Selected as key AI term
13	classifier	Removed
14	ontological	Removed
14	ontology	Removed
14	KNOWLEDGE_BASE	Key AI term
15	logistic_regression	Removed
15	regression	Removed
16	ARTIFICIAL_INTELLIGENCE	Key AI term
16	computational_intelligence	Selected as Key AI term
16	MACHINE_LEARNING	Key AI term
16	PATTERN_RECOGNITION	Key AI term
17	name_entity_recognition	Selected as key AI term
17	opinion_mining	Selected as key AI term

Cluster number	Quasi-synonyms or key AI terms	Status
17	text_categorization	Selected as key AI term
17	text_summarization	Selected as key AI term
17	word_sense_disambiguation	Selected as key AI term
18	fuzzy	Removed
19	markov_decision_process	Selected as key AI term
20	humanoid	Selected as key AI term
20	humanoid_robot	Removed
20	wheelchair	Removed
20	ROBOT	key AI term
21	IMAGE_ALIGNMENT	key AI term
21	camera_calibration	Selected as key AI term
21	mosaicing	Removed
21	non_rigid_registration	Selected as key AI term
21	registration	Removed
21	rigid_registration	Selected as key AI term
21	stereo_matching	Selected as key AI term
22	artificial_neural_network	Selected as key AI term
22	feed_forward_neural_network	Selected as key AI term
22	multilayer_neural_network	Selected as key AI term
22	neural_net	Selected as key AI term
22	perceptron	Selected as key AI term
22	recurrent_neural_network	Selected as key AI term
22	NEURAL_NETWORK	Key AI term
23	neuro_fuzzy	Selected as key AI term
23	radial_basis_function	Selected as key AI term
23	self_organizing_map	Selected as key AI term
24	analog_vlsi	Selected as key AI term
24	NEUROMORPHIC_COMPUTING	Key AI term
24	associative_memory	Selected as key AI term
24	neuromorphic	Removed
24	neuromorphic_hardware	Selected as key AI term
24	spike_neural_network	Selected as key AI term
25	back_propagation_neural_network	Selected as key AI term
25	bp_neural_network	Selected as key AI term
25	elman_network	Selected as key AI term
25	elman_neural_network	Selected as key AI term
25	less_square_support_vector_machine	Selected as key AI term
25	rbf_neural_network	Selected as key AI term
26	adaboost	Selected as key AI term
26	decision_tree	Selected as key AI term
26	random_forest	Selected as key AI term
26	ensemble	Removed
26	SUPPORT_VECTOR_MACHINE	Key AI term
27	KNOWLEDGE_REPRESENTATION_AND_REASONING	Key AI term
27	commonsense_reasoning	Selected as key AI term
27	description_logic	Selected as key AI term
27	nonmonotonic_reasoning	Selected as key AI term
27	reasoning	Removed
28	IMAGE_MATCHING	Key AI term
28	alignment	Removed
28	match	Removed
28	template_matching	Selected as key AI term
29	dimensionality_reduction	Selected as key AI term

Cluster number	Quasi-synonyms or key AI terms	Status
29	discriminant_analysis	Selected as key AI term
29	principal_component_analysis	Selected as key AI term
30	DEEP_LEARNING	Key AI term
30	convolutional_neural_network	Selected as key AI term
30	deep_belief_network	Selected as key AI term
30	deep_convolutional_neural_network	Selected as key AI term
30	deep_neural_network	Selected as key AI term
31	REINFORCEMENT_LEARNING	Key AI term
31	actor_critic	Selected as key AI term
31	sarsa	Selected as key AI term
32	COMPUTER_VISION	Key AI term
32	computer_graphic	Removed
32	machine_vision	Selected as key AI term
33	person_re_identification	Selected as key AI term
34	knowledge	Removed
35	back_propagation	Selected as key AI term
36	ROBOT_SYSTEM	Key AI term
36	manipulator	Removed
37	BAYESIAN_BELIEF_NETWORK	Key AI term
37	COLLABORATIVE_SYSTEM	Key AI term
37	MACHINE_INTELLIGENCE	Key AI term
37	PATTERN_ANALYSIS	Key AI term
37	SYMBOLIC_REASONING	Key AI term
37	SYSTEM_AND_CONTROL_THEORY	Key AI term
37	neuromolecular	Removed
38	NATURAL_LANGUAGE_PROCESSING	Key AI term
38	machine_translation	Selected as key AI term
38	question_answering	Selected as key AI term
39	linear_discriminant	Selected as key AI term
39	multiclass_classification	Selected as key AI term
39	rbf_kernel	Selected as key AI term
40	SENSOR_DATUM_FUSION	Key AI term

Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018

Table B.2. Selected list of Key AI term

Selected terms for document retrieval within the 13 funding databases

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
actor critic	actor critic	Scopus	non-core
adaboost	adaboost	Scopus	CORE
analog vlsi	analog vlsi	Scopus	non-core
artificial intelligence	artificial intelligence	M & C	CORE
artificial neural networks	artificial neural network	Scopus	CORE
associative memory	associative memory	Scopus	non-core
autonomous vehicle	autonomous vehicle	Funding databases	non-core
back propagation	back propagation	Scopus	CORE
back propagation neural network	back propagation neural network	Scopus	CORE
bagging	bagging	Funding databases	non-core
bayes classifier	bayes classifier	Scopus	non-core
bayesian belief networks	bayesian belief network	C	non-core
bioinformatics	bioinformatics	Scopus	non-core
bp neural network	bp neural network	Scopus	Merged to back propagation neural network
camera calibration	camera calibration	Scopus	non-core
collaborative systems	collaborative system	C	non-core
commonsense reasoning	commonsense reasoning	Scopus	non-core
computational biology	computational biology	Scopus	non-core
computational intelligence	computational intelligence	Scopus	CORE
computer vision	computer vision	C	CORE
convolutional neural network	convolutional neural network	Scopus	CORE
data mining	datum mining	Funding databases	non-core
decision tree	decision tree	Scopus	non-core
deep belief network	deep belief network	Scopus	CORE
deep convolutional neural network	deep convolutional neural network	Scopus	CORE
deep learning	deep learning	C	CORE
deep neural network	deep neural network	Scopus	CORE
description logic	description logic	Scopus	non-core
dimensionality reduction	dimensionality reduction	Scopus	non-core
discriminant analysis	discriminant analysis	Scopus	non-core
elman network	elman network	Scopus	CORE
elman neural network	elman neural network	Scopus	CORE
ensemble learning	ensemble learning	Funding databases	non-core
expert systems	expert system	M	CORE
feed forward neural network	fee forward neural network	Scopus	CORE
fuzzy logic	fuzzy logic	M	non-core
gene ontology	gene ontology	M	non-core
hidden markov model	hide markov model	NSF	non-core
humanoid	humanoid	Scopus	non-core
humanoid robotics	humanoid robotic	C	Merged to robot
image alignment	image alignment	C	non-core
image matching	image match	C	non-core

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
inference engine	inference engine	Scopus	CORE
information retrieval	information retrieval	Funding databases	non-core
k nearest neighbors	k near neighbor	Scopus	Merged to near neighbor classifier
k nn classifier	k nn classifier	Scopus	Merged to near neighbor classifier
kegg pathway	kegg pathway	Scopus	non-core
knowledge bases	knowledge base	M	non-core
knowledge representation and reasoning	knowledge representation and reasoning	C	non-core
least square support vector machines	less square support vector machine	Scopus	Merged to support vector machine
linear discriminant	linear discriminant	Scopus	non-core
machine intelligence	machine intelligence	C	CORE
machine learning	machine learning	M & C	CORE
machine translation	machine translation	Scopus	CORE
machine vision	machine vision	Scopus	CORE
markov decision process	markov decision process	Scopus	non-core
multiclass classification	multiclass classification	Scopus	non-core
multilayer neural network	multilayer neural network	Scopus	CORE
naive bayes	naive bayes	Scopus	non-core
name entity recognition	name entity recognition	Scopus	non-core
natural language processing	natural language processing	M & C	CORE
nearest neighbor classifier	near neighbor classifier	Scopus	non-core
neural net	neural net	Scopus	Merged to neural network
neural networks	neural network	M & C	non-core
neuro fuzzy	neuro fuzzy	Scopus	non-core
neuromorphic computing	neuromorphic computing	C	non-core
neuromorphic hardware	neuromorphic hardware	Scopus	non-core
non rigid registration	non rigid registration	Scopus	non-core
nonmonotonic reasoning	nonmonotonic reasoning	Scopus	non-core
object recognition	object recognition	Funding databases	non-core
opinion mining	opinion mining	Scopus	non-core
optimal search	optimal search	C	non-core
pattern analysis	pattern analysis	C	non-core
pattern recognition	pattern recognition	C	non-core
perceptron	perceptron	Scopus	CORE
person re identification	person re identification	Scopus	non-core
principal component analysis	principal component analysis	Scopus	non-core
question answering	question answering	Scopus	non-core
radial basis function	radial basis function	Scopus	non-core
random forest	random forest	Scopus	CORE
rbf kernel	rbf kernel	Scopus	non-core
rbf neural network	rbf neural network	Scopus	CORE
recurrent neural network	recurrent neural network	Scopus	CORE
reinforcement learning	reinforcement learning	C	non-core
rigid registration	rigid registration	Scopus	non-core
robot systems	robot system	C	Merged to robot
robotics	robotic	M & C	Merged to robot
robots	robot	M & C	non-core
sarsa	sarsa	Scopus	non-core

Terms before lemmatisation	Terms after lemmatisation	Source	Term Status
self organizing map	self organizing map	Scopus	CORE
sensor data fusion	sensor datum fusion	C	non-core
sensor networks	sensor network	C	non-core
speech recognition	speech recognition	Funding databases	non-core
spike neural network	spike neural network	Scopus	CORE
stereo matching	stereo matching	Scopus	non-core
supervised learning	supervised learning	M	CORE
support vector machines	support vector machine	M	CORE
svm classifier	svm classifier	Scopus	CORE
symbolic reasoning	symbolic reasoning	C	non-core
systems and control theory	system and control theory	C	non-core
template matching	template matching	Scopus	non-core
text categorization	text categorization	Scopus	non-core
text mining	text mining	Funding databases	non-core
text summarization	text summarization	Scopus	non-core
unsupervised learning	unsupervised learning	M & C	CORE
word sense disambiguation	word sense disambiguation	Scopus	non-core

Note: In the "Source" column, "M" refers to the MeSH (M-list) and "C" refers to the list in Cockburn et al. (2018) (C-list), "Scopus" refers to terms retrieved from Scopus that are "similar" to core terms in M-C lists, and "Funding databases" refer to quasi-synonyms retrieved from some of the 13 databases analysed. The column "Term Status" refers to how each term was treated for analysis and retrieval in each corresponding database. "CORE" indicates that the term was used as a core AI term and was not penalised for potential ambiguity, while "non-core" indicates that the term was used but was partly penalising when deciding which documents to select.
Source: OECD calculations based on Scopus Custom Data, Elsevier, Version 1.2018; on project microdata and results provided by the Netherlands and Spain.

Annex C. Results by database and agency

This section presents the detailed funding allocations of the 13 funding databases analysed. This supplements the results shown in section 3 of the main text. A table for year-by-year allocations and a figure illustrating the trends in project numbers and funding amounts are provided for each database.

ARC (AUS) funding

Table C.1 shows that the number of projects identified as AI-related has fluctuated yet increased from 15 in 2002 to around 30 in 2018 and 2019. The amount of R&D funding displays a similar trend, fluctuating between USD 1 million and USD 16 million. Note that the abstracts in ARC are mostly brief (only 3 to 5 sentences) and do not necessarily contain much information on the nature of the projects; this may affect the results of the analysis.

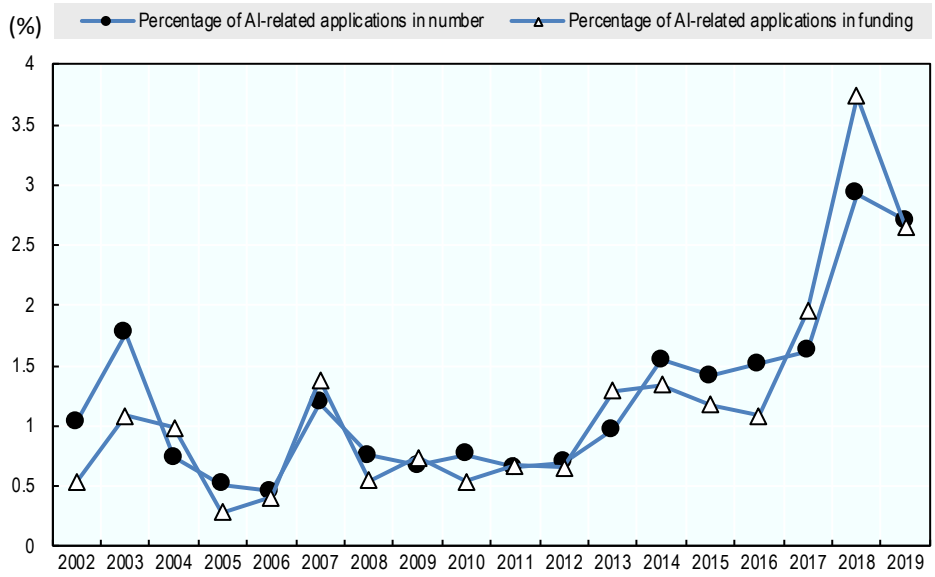
Table C.1. Estimates of AI-related R&D in ARC funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2002	15	1 454	1.03	1	275	0.53
2003	34	1 917	1.77	8	737	1.08
2004	12	1 633	0.73	4	356	0.98
2005	9	1 753	0.51	2	526	0.29
2006	7	1 530	0.46	1	365	0.41
2007	17	1 428	1.19	5	361	1.38
2008	11	1 459	0.75	2	386	0.55
2009	11	1 637	0.67	4	503	0.74
2010	13	1 697	0.77	3	578	0.54
2011	11	1 679	0.66	5	739	0.67
2012	12	1 719	0.70	4	569	0.65
2013	15	1 552	0.97	7	538	1.29
2014	22	1 421	1.55	11	791	1.34
2015	18	1 268	1.42	5	426	1.18
2016	19	1 256	1.51	5	445	1.09
2017	18	1 110	1.62	12	632	1.95
2018	33	1 127	2.93	16	437	3.74
2019	28	1 037	2.70	9	329	2.66

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on ARC Grants Search data, accessed August 2020.

Figure C.1. Estimates of AI-related ARC funding

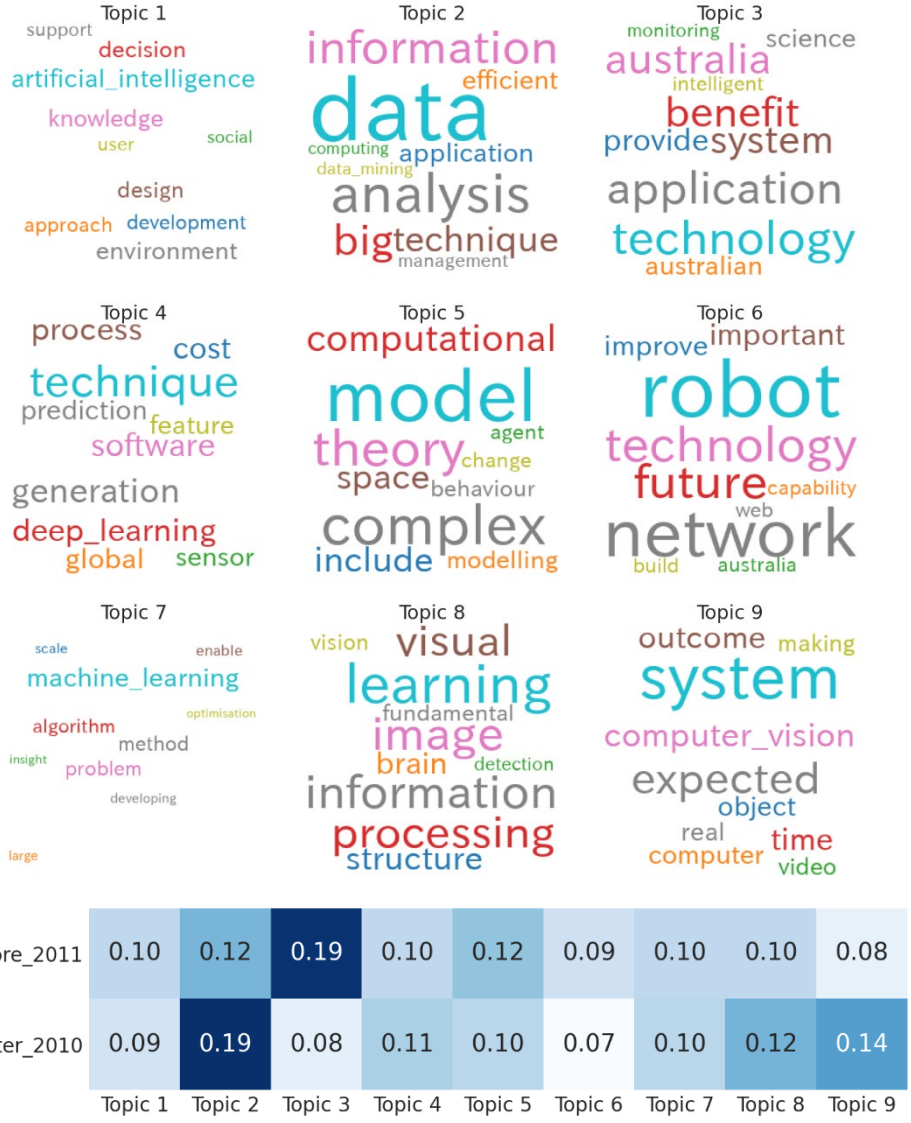
As a percentage of total ARC funding (number of projects and funding amounts)



Source: OECD calculations based on ARC Grants Search data, accessed August 2020.

Figure C.2 shows the topics identified by applying the LDA algorithm to the documents associated with the AI-related R&D projects funded by the ARC. The topics retrieved were 1. AI for decision support, 2. Big data analysis, 3. Technology for intelligent system(s), 4. Deep learning technique(s), 5. Complex model(s), 6. Robot network(s), 7. Machine learning analysis, 8. Image processing, and 9. Computer vision.

Figure C.2. Topics from the AI-related documents of the ARC with relative topic prominence in different periods, 2002-2019



Note: The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on ARC Grants Search data, accessed August 2020.

CIHR (CAN) funding

Table C.2 shows that the number of projects identified as AI-related was fewer than 10 per year until 2013, when that number began to increase, reaching 49 in 2018. Yearly R&D funding amounts held stable at around USD 1 million until surging in 2016. **Figure C.3** shows upward trends after 2014 and a surge in 2016 in both the percentage of funded projects that were related to AI and in the percentage of funding AI-related projects received.

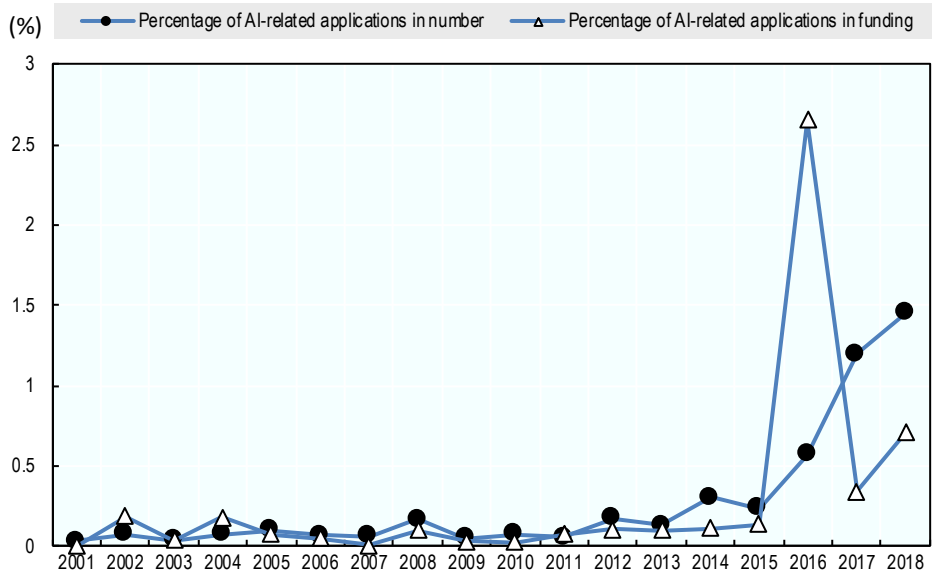
Table C.2. Estimates of AI-related R&D in CIHR funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	1	2 925	0.03	0	638	0.00
2002	2	2 684	0.07	1	553	0.19
2003	1	2 713	0.04	0	671	0.04
2004	2	2 529	0.08	1	595	0.19
2005	3	2 961	0.10	1	687	0.08
2006	2	2 787	0.07	0	724	0.05
2007	2	2 966	0.07	0	795	0.01
2008	6	3 561	0.17	1	861	0.10
2009	2	4 000	0.05	0	826	0.03
2010	3	3 877	0.08	0	943	0.02
2011	2	3 478	0.06	1	812	0.08
2012	6	3 429	0.17	1	676	0.11
2013	4	3 046	0.13	1	690	0.10
2014	10	3 259	0.31	1	913	0.11
2015	7	2 955	0.24	1	974	0.14
2016	18	3 150	0.57	27	1 008	2.66
2017	37	3 084	1.20	3	824	0.34
2018	49	3 374	1.45	7	959	0.71

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on CIHR Canadian Research Information System data, accessed August 2020.

Figure C.3. Estimates of AI-related CIHR funding

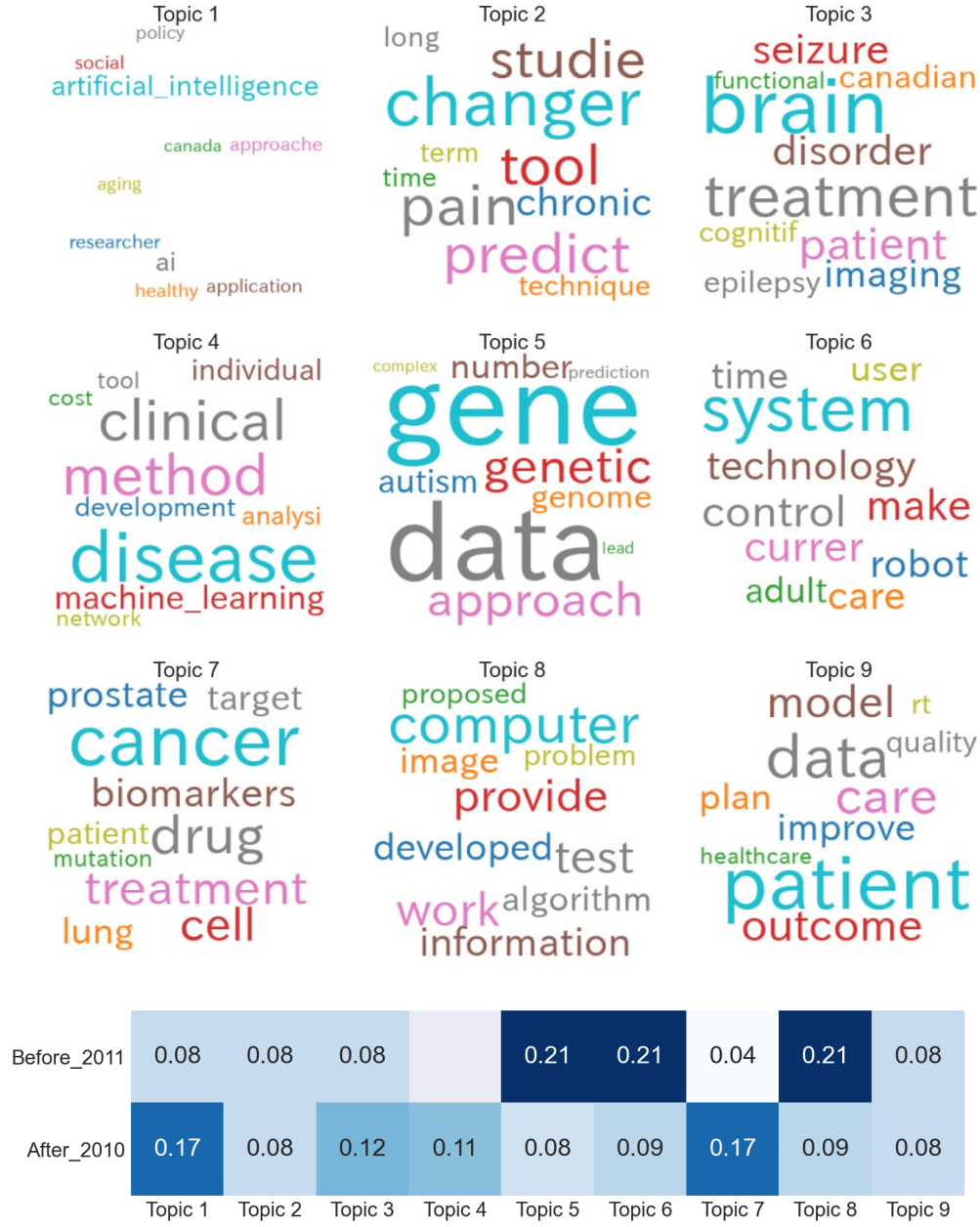
As a percentage of total CIHR funding (number of projects and funding amounts)



Source: OECD calculations based on CIHR Canadian Research Information System data, accessed August 2020.

Figure C.4 shows the topics identified from the documents associated with the AI-related R&D projects funded by CIHR. The topics retrieved were 1. AI application(s) for society, 2. Studies related to chronic pain, 3. Brain disorder and treatment, 4. Machine learning for diagnosis, 5. Genetic data analysis, 6. Health robot system(s), 7. Cancer treatment, 8. Image analysis, and 9. Improving patient care.

Figure C.4. Topics from the AI-related documents of the CIHR with relative topic prominence in different periods, 2001-2018



Note: The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on CIHR (Canadian Research Information System) data, accessed August 2020.

NSERC (CAN) funding

Table C.3 shows that the number of projects identified as AI-related increased by more than eighteenfold from 21 in 2001 to 384 in 2017. The amount of R&D funding displays a similar trend, having increased from USD 1 million in 2001 to USD 10 million in 2017. The number of AI-related projects has increased from 0.3% to 3.2% of total funded projects; likewise, the amount of funding for AI-related projects increased from 0.4% to 4.4% of total funding, as shown in **Figure C.5**.

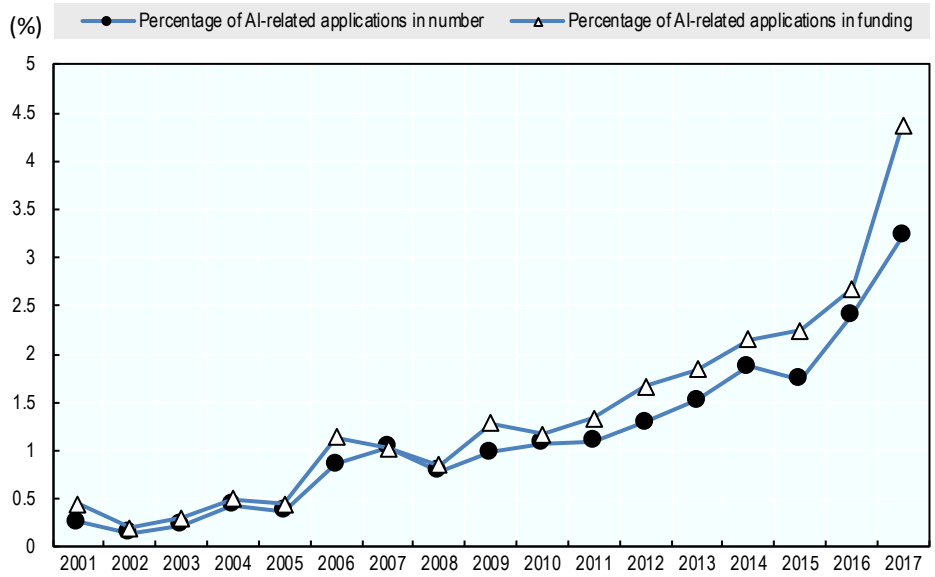
Table C.3. Estimates of AI-related R&D in NSERC funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	21	7 900	0.27	1	135	0.44
2002	12	8 335	0.14	0	159	0.20
2003	22	9 370	0.23	0	159	0.31
2004	42	9 678	0.43	1	165	0.50
2005	40	10 551	0.38	1	187	0.45
2006	87	10 071	0.86	2	168	1.14
2007	113	10 879	1.04	3	259	1.03
2008	94	11 987	0.78	2	224	0.85
2009	105	10 679	0.98	3	226	1.28
2010	112	10 395	1.08	3	214	1.18
2011	105	9 531	1.10	2	187	1.33
2012	131	10 151	1.29	3	201	1.66
2013	162	10 680	1.52	4	204	1.84
2014	209	11 169	1.87	4	209	2.15
2015	196	11 291	1.74	5	227	2.24
2016	274	11 353	2.41	7	250	2.68
2017	384	11 925	3.22	10	230	4.38

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NSERC's Awards Database, accessed August 2020.

Figure C.5. Estimates of AI-related NSERC funding

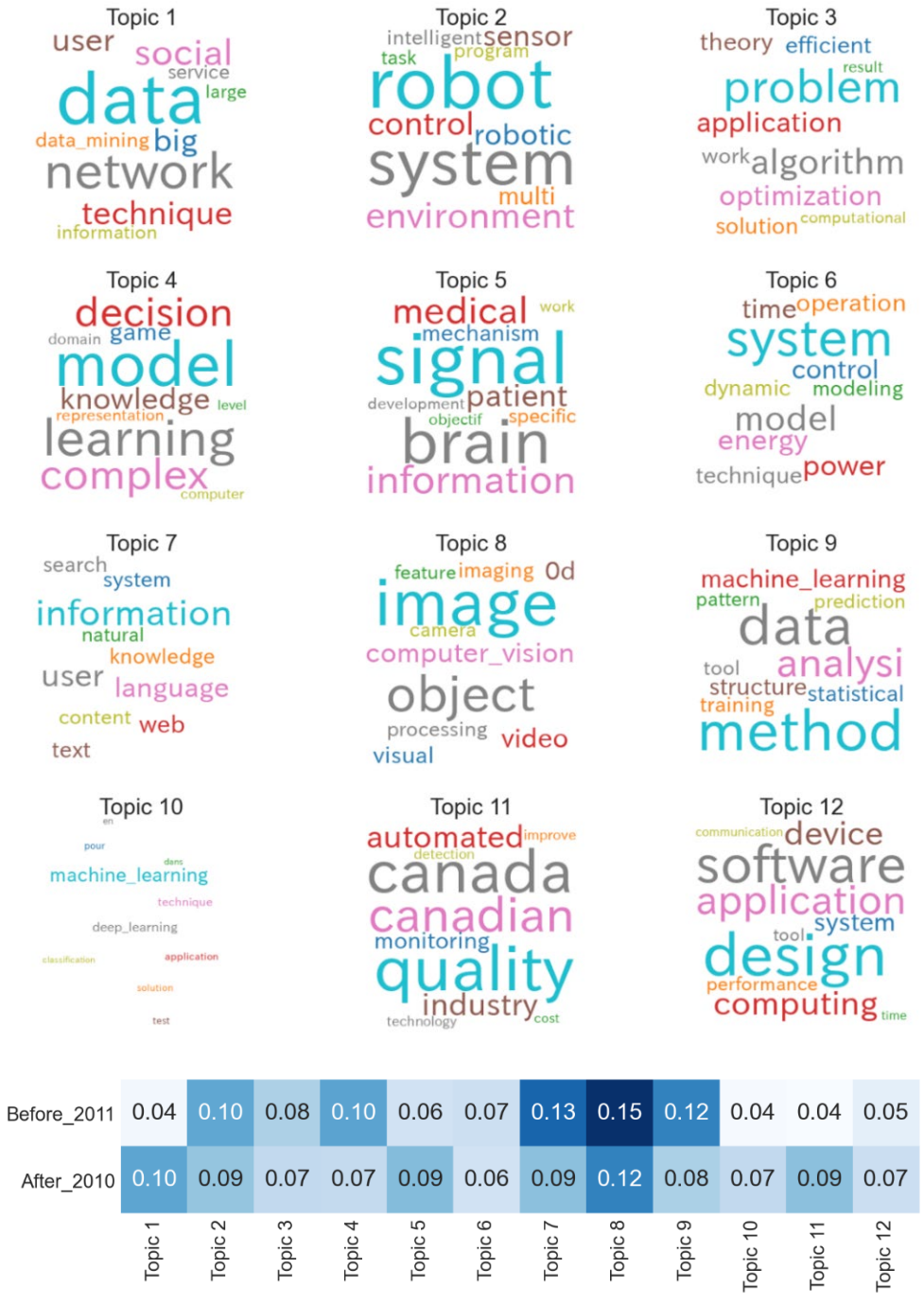
As a percentage of total NSERC funding (number of projects and funding amounts)



Source: OECD calculations based on NSERC's Awards Database, accessed August 2020.

Figure C.6 shows the topics identified from the documents associated with the AI-related R&D projects funded by NSERC. The topics retrieved were 1. Big data analysis, 2. Robot system(s), 3. Optimisation algorithm(s), 4. Decision support, 5. Brain analysis, 6. Energy system(s), 7. Natural language processing, 8. Computer vision, 9. Machine learning based data analysis, 10. Machine learning and deep learning, 11. Quality Canadian industry, and 12. Software design.

Figure C.6. Topics from the AI-related documents of NSERC with relative topic prominence in different periods, 2001-2017



Note: The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on NSERC's Awards Database, accessed August 2020.

PlanEst (ESP) funding

Table C.4 shows that the number of projects identified as AI-related has fluctuated between 69 and 187 throughout the period from 2004 to 2016. The amount of yearly R&D funding displays a similar trend, having fluctuated between USD 10 million and USD 56 million. **Figure C.7** illustrates the fluctuations in AI-project funding as percentages of the total yearly number of projects and amount of funding.

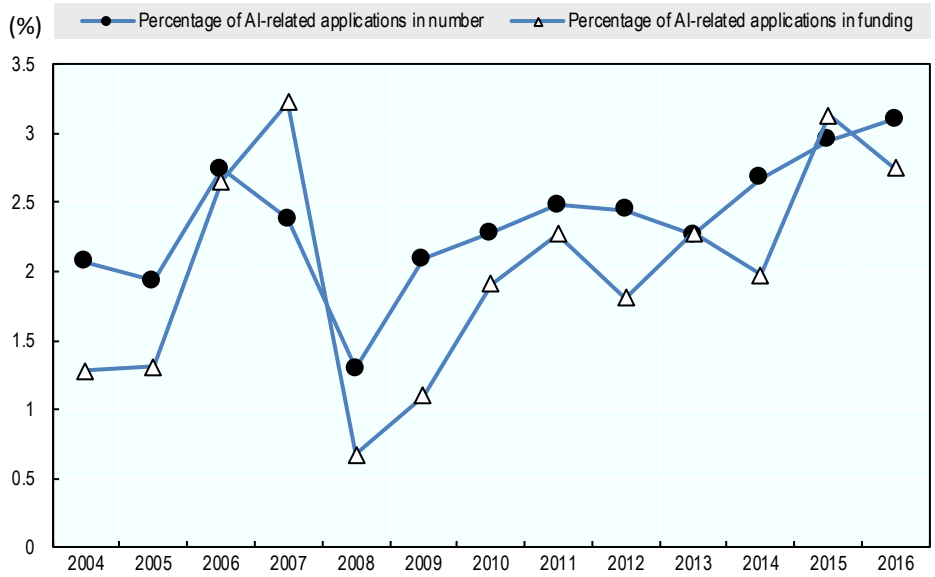
Table C.4. Estimates of AI-related R&D in PlanEst funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2004	80	3 868	2.07	10	790	1.28
2005	79	4 094	1.93	12	915	1.31
2006	129	4 710	2.74	38	1 419	2.66
2007	105	4 409	2.38	56	1 748	3.23
2008	69	5 333	1.29	14	2 027	0.67
2009	133	6 355	2.09	32	2 945	1.10
2010	135	5 922	2.28	44	2 311	1.91
2011	145	5 844	2.48	55	2 406	2.27
2012	118	4 822	2.45	24	1 304	1.81
2013	132	5 828	2.26	39	1 702	2.28
2014	145	5 417	2.68	33	1 685	1.97
2015	187	6 339	2.95	52	1 668	3.13
2016	150	4 829	3.11	37	1 336	2.75

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: FECYT calculations by the codes developed by the OECD based on PlanEst data, accessed July 2017.

Figure C.7. Estimates of AI-related PlanEst funding

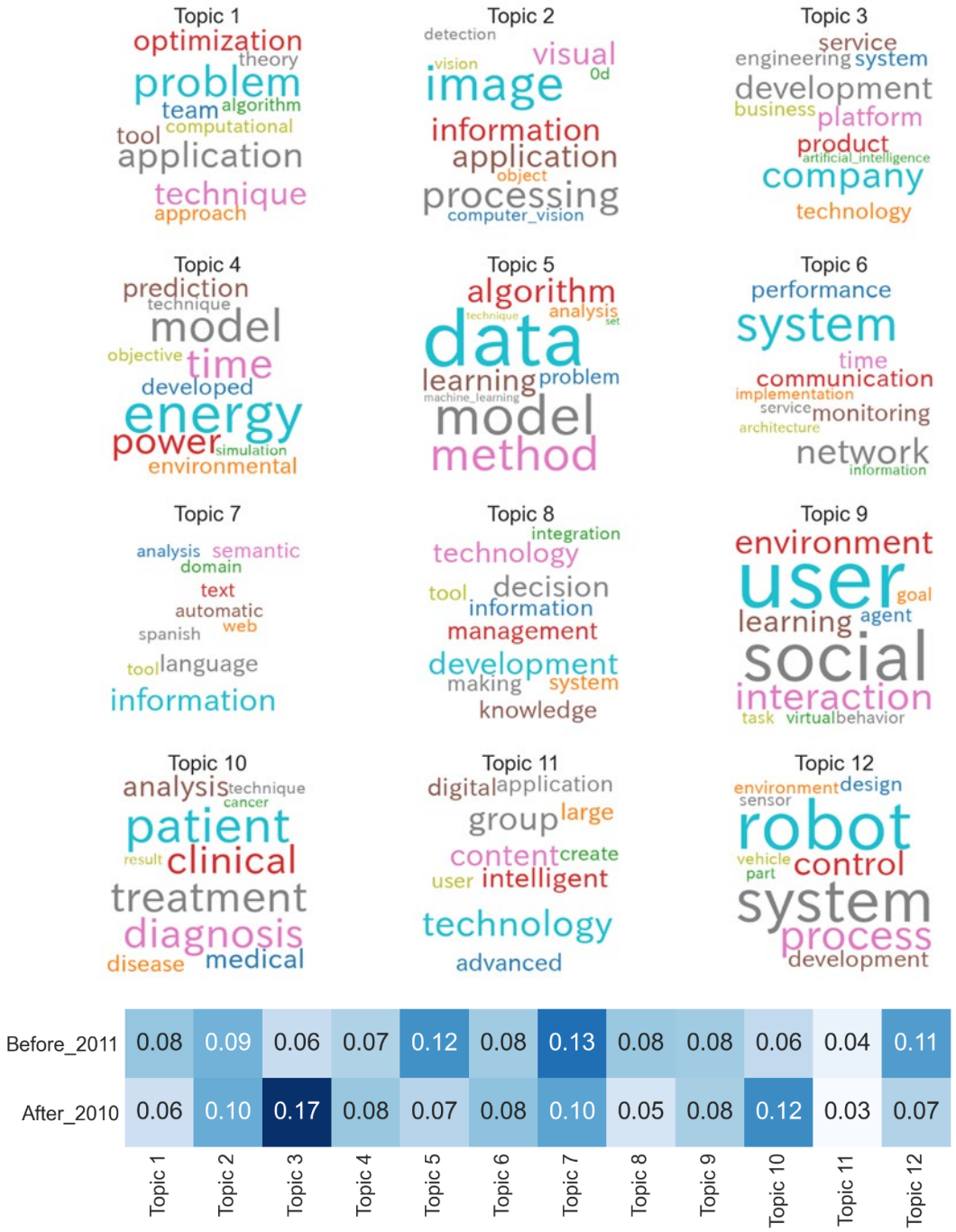
As a percentage of total PlanEst funding (number of projects and funding amounts)



Source: FECYT calculations by the codes developed by the OECD based on PlanEst data, accessed July 2017.

Figure C.8 shows the topics identified from the documents associated with the AI-related R&D projects funding by the PlanEst. The topics retrieved were 1. Optimisation algorithms, 2. Image processing, 3. AI application(s) to business, 4. Energy-related modelling, 5. Machine learning algorithms, 6. Network system(s), 7. Natural language processing, 8. Decision support, 9. Brain analysis, 10. Clinical analysis, 11. Intelligent technology, and 12. Robot system(s).

Figure C.8. Topics from the AI-related documents of the PlanEst with relative topic prominence in different periods, 2004-2016



Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: FECYT calculations by the codes developed by the OECD based on PlanEst data, accessed July 2017.

ANR (FRA) funding

Table C.5 shows that the number of projects identified as AI-related fluctuated until 2013 and then increased, reaching 57 in 2019. The amount of R&D funding displays a similar trend, having increased to USD 27 million in 2019 after having fluctuated until 2013. Both the percentage of AI-related projects and the percentage of funding allocated to AI-related projects increased from 0.2% in 2013 to around 4.1% in 2019, as shown in **Figure C.9**.

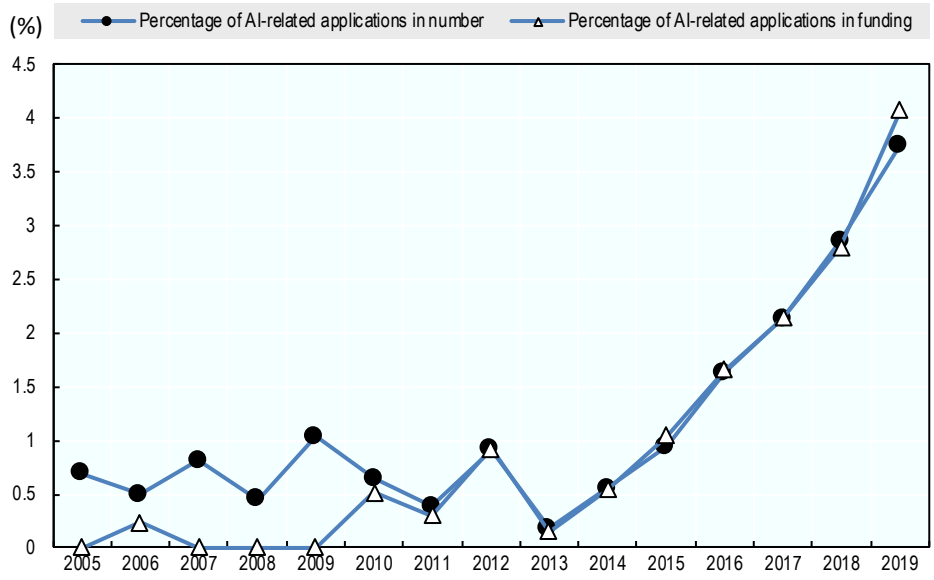
Table C.5. Estimates of AI-related R&D in ANR funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2005	10	1 441	0.69	0	15	0.00
2006	8	1 605	0.50	0	51	0.23
2007	12	1 478	0.81	0	119	0.00
2008	6	1 335	0.45	0	164	0.00
2009	14	1 355	1.03	0	161	0.00
2010	9	1 387	0.65	4	705	0.52
2011	5	1 312	0.38	2	632	0.30
2012	12	1 306	0.92	6	649	0.92
2013	2	1 110	0.18	1	517	0.16
2014	6	1 080	0.56	3	495	0.54
2015	10	1 057	0.95	5	466	1.04
2016	21	1 291	1.63	9	553	1.65
2017	30	1 408	2.13	14	636	2.14
2018	41	1 436	2.86	19	683	2.80
2019	57	1 522	3.75	27	659	4.07

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on Appels à projects ANR, accessed August 2020.

Figure C.9. Estimates of AI-related ANR funding

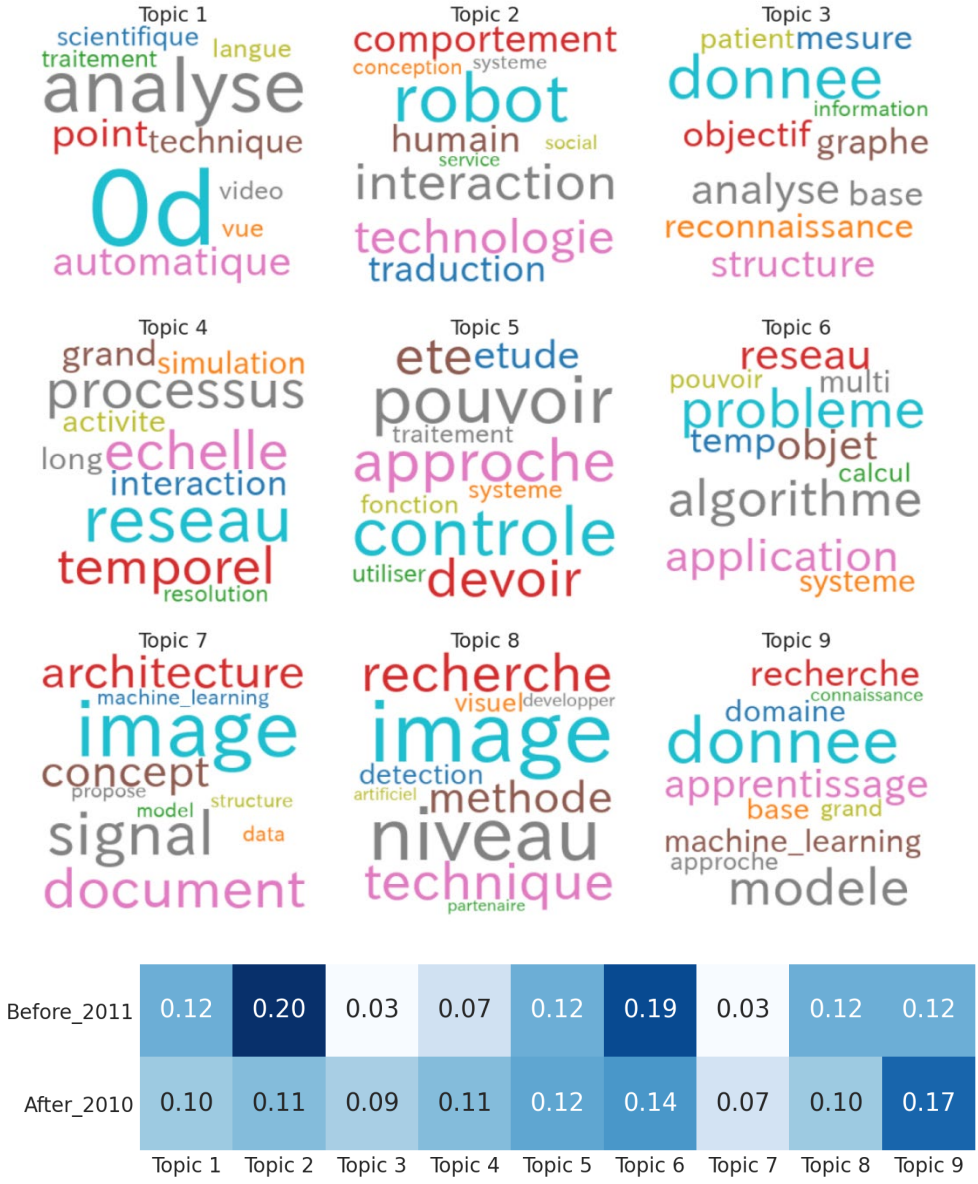
As a percentage of total ANR funding (number of projects and funding amounts)



Source: OECD calculations based on Appels à projets ANR, accessed August 2020.

Figure C.10 shows the topics identified from the documents associated with the AI-related R&D projects funded by the ANR. The topics retrieved were 1. 3D image analysis (numbers are all converted into 0 in the cleaning process, which results in producing "0d"), 2. Human robot interaction, 3. Patient data analysis, 4. Realtime network(s), 5. Treatment improvement, 6. Algorithm application(s), 7. Machine learning for image(s), 8. Image analysis technique(s), and 9. General machine learning.

Figure C.10. Topics from the AI-related documents of the ANR with relative topic prominence in different periods, 2005-2019



Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on Appels à projets ANR, accessed August 2020.

GtR (Innovate UK, GBR) funding

Table C.6 shows that the number of projects identified as AI-related increased nearly seventyfold from three in 2008 to 206 in 2019, surging after 2016. The amount of R&D funding displays a similar trend, having increased from USD 2 million in 2008 to USD 181 million in 2019. The percentage of AI-related projects increased from 1.0% to 13.3%, while the percentage of funding allocated to AI-related projects increased from 0.6% to 10.8% despite briefly dropping to 2.7% in 2018, as shown in **Figure C.11.**

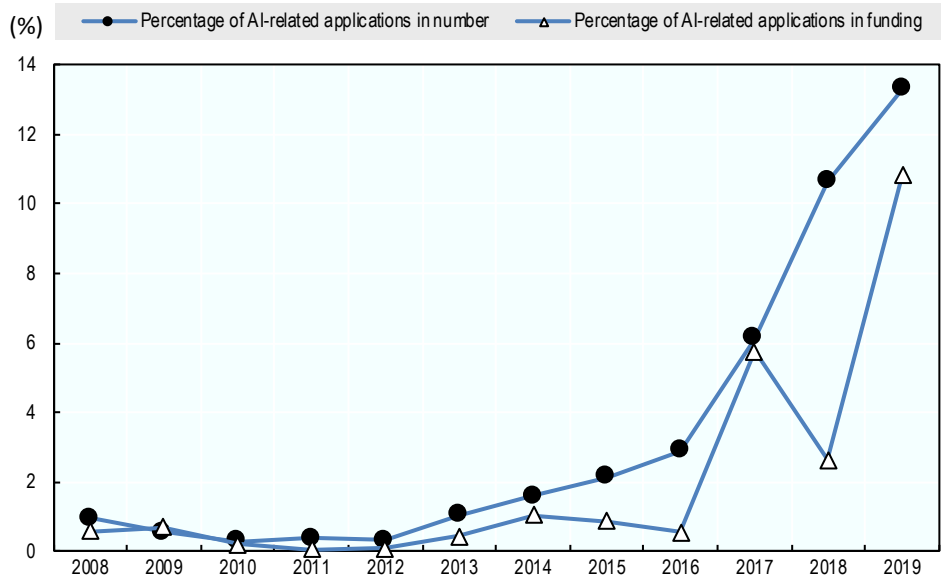
Table C.6. Estimates of AI-related R&D in GtR (Innovate UK part) funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2008	3	314	0.96	2	333	0.60
2009	3	545	0.55	3	439	0.69
2010	2	669	0.30	1	419	0.22
2011	4	1 030	0.39	0	525	0.07
2012	5	1 545	0.32	1	791	0.09
2013	26	2 458	1.06	4	883	0.46
2014	39	2 440	1.60	13	1 197	1.06
2015	67	3 097	2.16	13	1 501	0.87
2016	38	1 304	2.91	9	1 533	0.56
2017	118	1 924	6.13	90	1 569	5.75
2018	165	1 550	10.65	91	3 425	2.65
2019	206	1 548	13.31	181	1 667	10.83

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on Gateway to Research (Innovate UK part), accessed August 2020.

Figure C.11. Estimates of AI-related GtR (Innovate UK part) funding

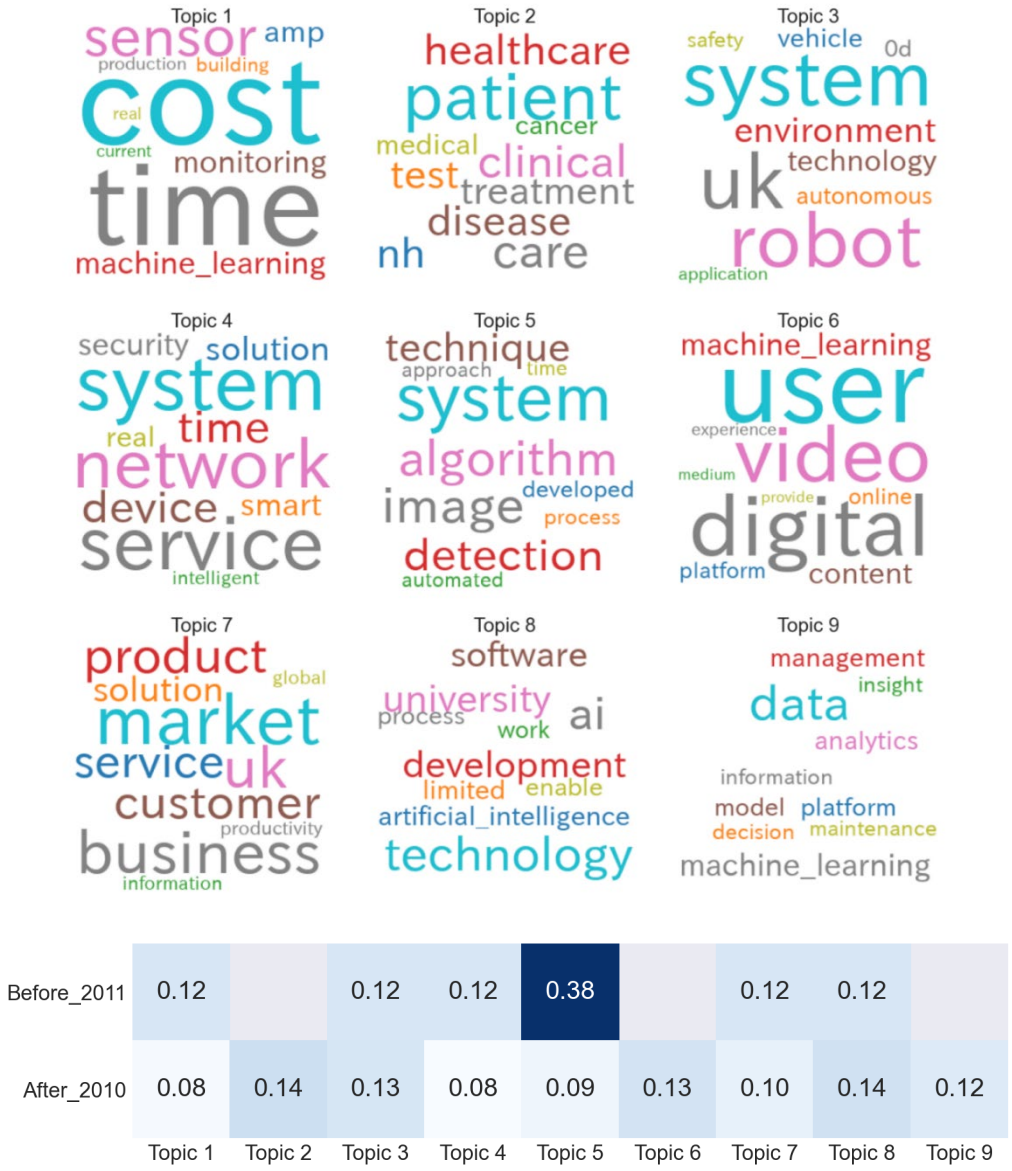
As a percentage of total GtR (Innovate UK part) funding (number of projects and funding amounts)



Source: OECD calculations based on Gateway to Research (Innovate UK component), accessed August 2020.

Figure C.12 shows the topics identified from the documents associated with the AI-related R&D projects funded by the Innovate UK component of the GtR. The topics retrieved were 1. Sensor monitoring, 2. Clinical care, 3. Robot system(s), 4. Service system(s), 5. Image detection system(s), 6. Machine learning based video analysis, 7. Market solution(s), 8. AI technology development, and 9. Decision support.

Figure C.12. Topics from the AI-related documents of the GtR-Innovate UK with relative topic prominence in different periods, 2008-2019



Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on Gateway to Research (Innovate UK part), accessed August 2020.

GtR (Research Councils, GBR) funding

Table C.7 shows that the number of projects identified as AI-related has increased nearly tenfold from 74 in 2006 to 701 in 2019, surging after 2015. The amount of R&D funding displays a similar trend, having increased from USD 30 million in 2006 to USD 552 million in 2019. The percentage of AI-related projects increased from 1.5% to 7.9%, while the percentage of funding allocated to AI-related projects increased from 1.1% to 13.5%, as shown in **Figure C.13.**

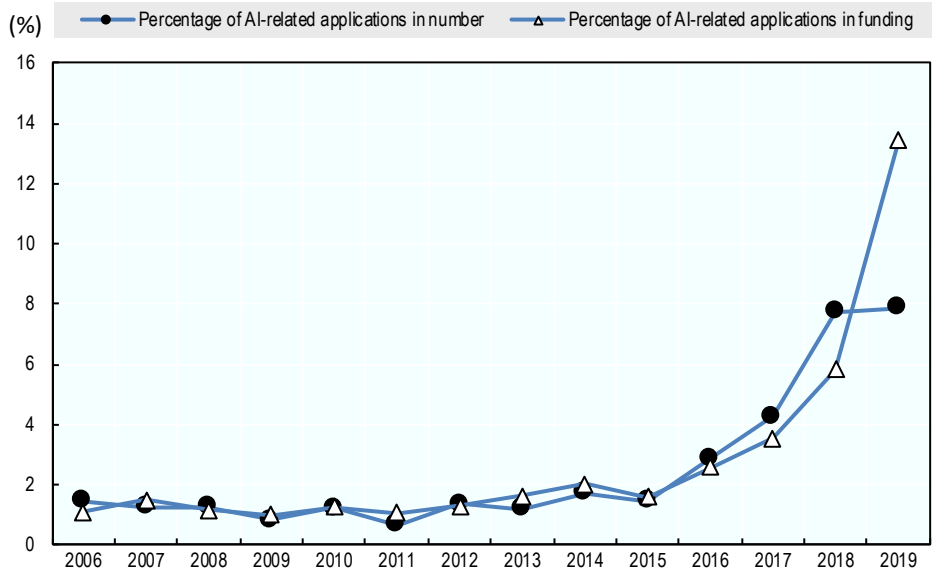
Table C.7. Estimates of AI-related R&D in GtR (Research Councils part) funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2006	74	5 063	1.46	30	2 701	1.10
2007	66	5 216	1.27	40	2 684	1.49
2008	65	5 176	1.26	35	3 058	1.16
2009	42	5 008	0.84	33	3 322	1.00
2010	66	5 286	1.25	37	2 945	1.26
2011	30	4 354	0.69	32	2 974	1.07
2012	63	4 603	1.37	38	2 955	1.29
2013	55	4 598	1.20	56	3 400	1.63
2014	73	4 198	1.74	80	3 946	2.02
2015	75	5 058	1.48	54	3 321	1.61
2016	182	6 271	2.90	88	3 404	2.58
2017	374	8 747	4.28	127	3 592	3.54
2018	639	8 259	7.74	226	3 879	5.84
2019	701	8 899	7.88	552	4 098	13.46

Note: These figures are results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on Gateway to Research (Research Councils component), accessed August 2020.

Figure C.13. Estimates of AI-related GtR (Research Councils part) funding

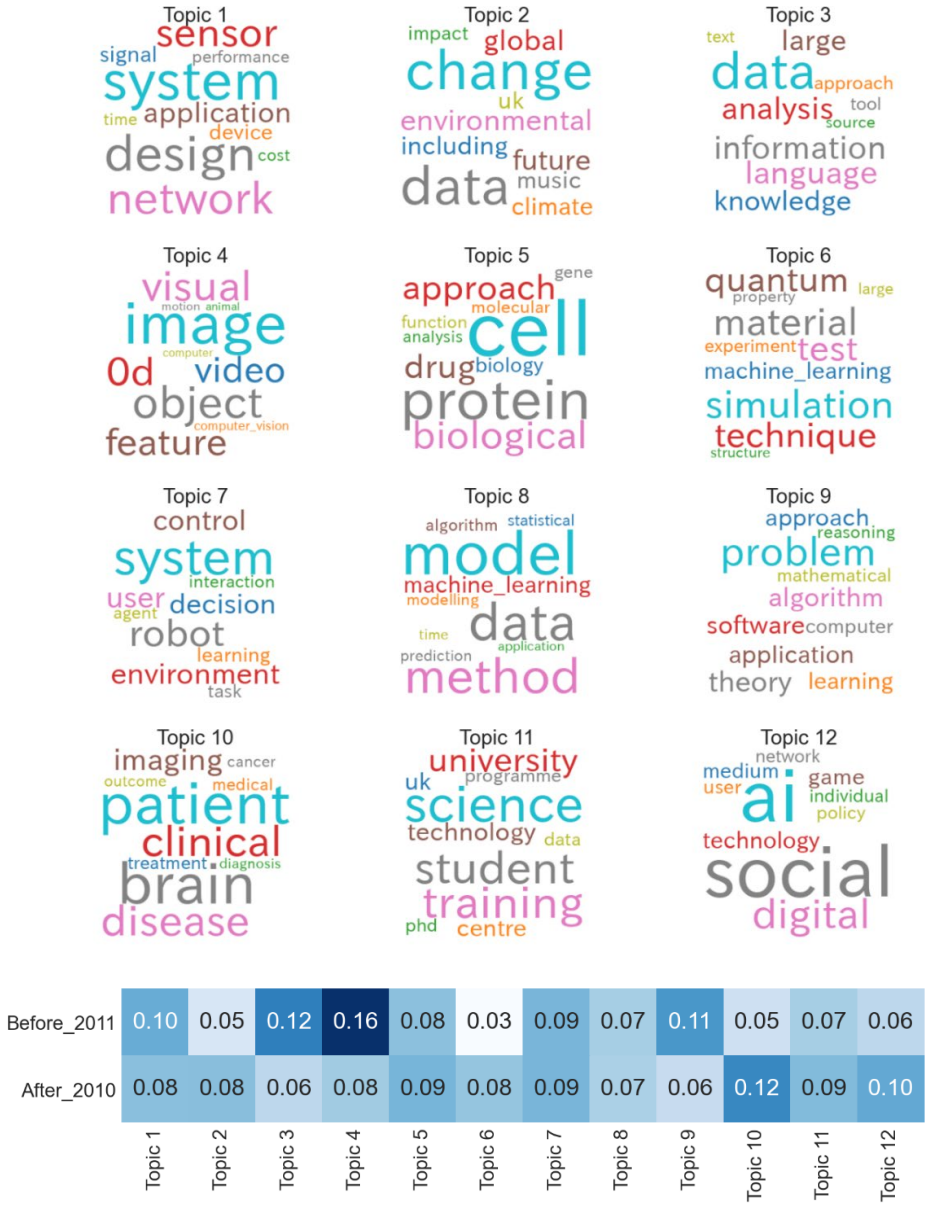
As a percentage of total GtR (Research Councils part) funding (number of projects and funding amounts)



Source: OECD calculations based on Gateway to Research (Research Councils part), accessed August 2020.

Figure C.14 shows the topics identified from the documents associated with the AI-related R&D projects funded by the Research Councils part of the GtR. The topics retrieved were 1. Sensor network system(s), 2. Climate change, 3. Language and knowledge analysis, 4. 3d image analysis, 5. Cell and protein, 6. Machine learning for materials, 7. Robot system(s), 8. General machine leaning, 9. Learning theory and algorithm(s), 10. Brain treatment, 11. Student training, and 12. AI for society.

Figure C.14. Topics from the AI-related documents of the GtR-Research Councils with relative topic prominence in different periods, 2006-2019



Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.

Source: OECD calculations based on Gateway to Research (Research Councils part), accessed August 2020.

AMED (JPN) funding

Table C.8 shows that the number of projects identified as AI-related increased fivefold from four in 2015 to 20 in 2018. The amount of R&D funding also increased from USD 4 million in 2015 to USD 13 million in 2018, surging in 2016 to USD 25 million. The percentage of AI-related projects increased from 0.2% to 2.5%, while the percentage of funding allocated to AI-related projects increased from 0.2% to 4.2%, as shown in **Figure C.15**.

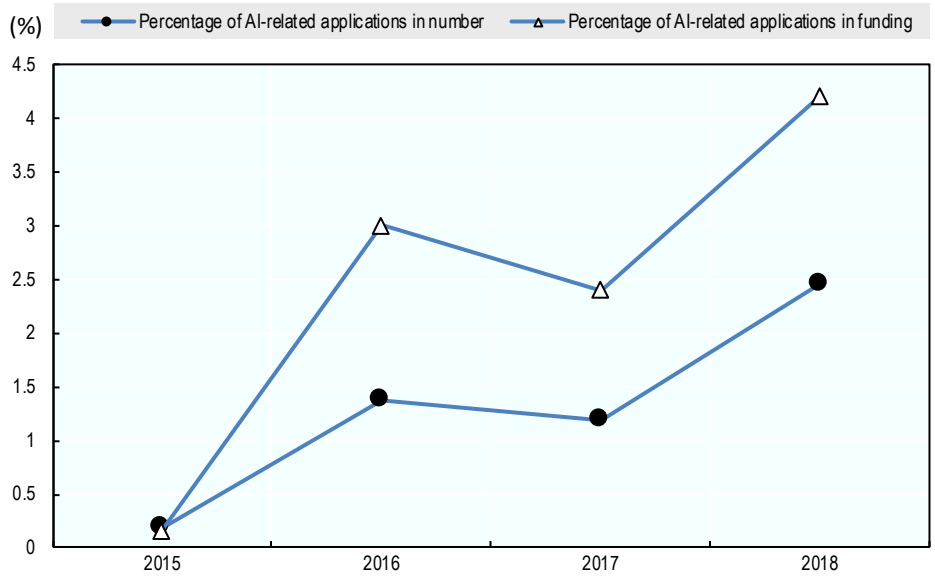
Table C.8. Estimates of AI-related R&D in AMED funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2015	4	2 072	0.19	4	2 287	0.16
2016	12	871	1.38	25	839	3.00
2017	12	1 009	1.19	19	786	2.40
2018	20	813	2.46	13	301	4.21

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on AMEDfind, accessed March 2020.

Figure C.15. Estimates of AI-related AMED funding

As a percentage of total AMED funding (number of projects and funding amounts)



Source: OECD calculations based on AMEDfind, accessed March 2020.

Figure C.16 shows the topics identified from the documents associated with the AI-related R&D projects funded by AMED. The topics retrieved were 1. Gene analysis, 2. Machine learning on cell analysis, 3. Medical research and analysis, 4. Data analysis for treatment, 5. Technological development related to genome, 6. Diagnosis support, 7. Neural circuit function, 8. AI system(s) on image(s), and 9. Diseases and patients.

Figure C.16. Topics from the AI-related documents of AMED with relative topic prominence, 2015-2018

Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on AMEDfind, accessed March 2020.

KAKEN (JPN) funding

Table C.9 shows the number of projects identified as AI-related slightly decreased or remained stable until 2008 before increasing nearly ninefold from 89 in 2008 to 788 in 2018. The amount of R&D funding shows a similar trend, fluctuating between USD 5 and 18 million until 2012 and then increasing to USD 70 million in 2017 and USD 62 million in 2018. The percentage of AI-related projects increased from around 0.5% to 2.7%, while the percentage of funding allocated to AI-related projects increased from around 0.5% to 2.9%, as shown in **Figure C.17**.

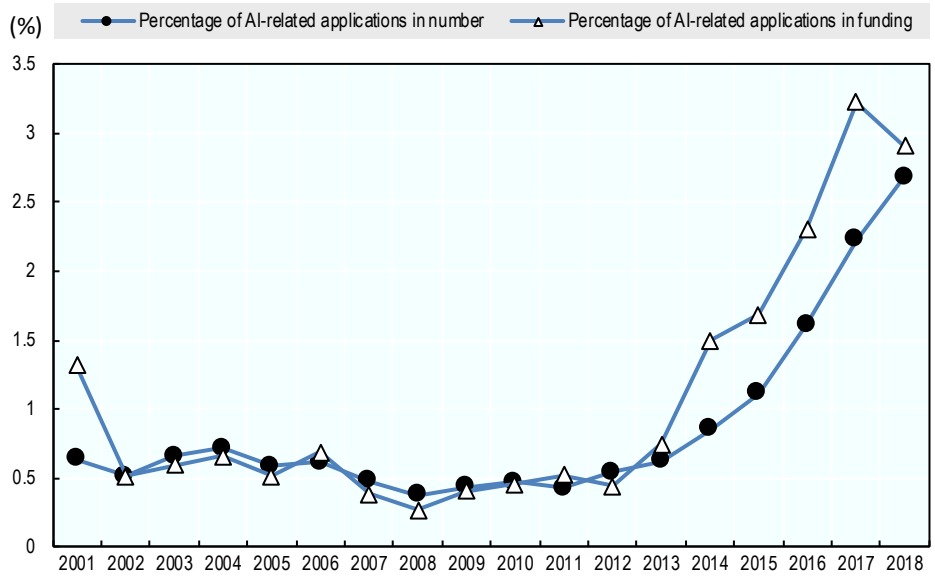
Table C.9. Estimates of AI-related R&D in KAKEN funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	134	21 034	0.64	18	1 374	1.32
2002	108	20 873	0.52	8	1 511	0.52
2003	141	21 385	0.66	9	1 443	0.60
2004	154	21 401	0.72	10	1 491	0.66
2005	138	23 386	0.59	10	1 910	0.51
2006	145	23 374	0.62	12	1 710	0.69
2007	116	23 988	0.48	7	1 736	0.38
2008	89	23 605	0.38	5	1 827	0.27
2009	112	25 447	0.44	8	1 889	0.41
2010	115	24 503	0.47	9	1 919	0.46
2011	129	29 961	0.43	11	2 181	0.53
2012	158	28 839	0.55	9	2 088	0.45
2013	183	29 440	0.62	16	2 127	0.75
2014	255	29 675	0.86	30	2 018	1.50
2015	339	30 375	1.12	35	2 098	1.69
2016	501	31 049	1.61	49	2 119	2.31
2017	646	28 998	2.23	70	2 182	3.22
2018	788	29 376	2.68	62	2 126	2.91

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on KAKEN, accessed August 2020.

Figure C.17. Estimates of AI-related KAKEN funding

As a percentage of total KAKEN funding (number of projects and funding amounts)



Source: OECD calculations based on KAKEN, accessed August 2020.

Figure C.18 shows the topics identified from the documents associated with the AI-related R&D projects funded by KAKEN. The topics retrieved were 1. Algorithm optimisation 2. Machine learning application(s) to cell and gene analysis, 3. International conference participation, 4. Robot control and learning, 5. Medical data analysis, 6 Image feature recognition, 7. Brain analysis, 8. Analytical system development, and 9. Natural language processing.

Figure C.18. Topics from the AI-related documents of KAKEN with relative topic prominence in different periods, 2001-2018

Note: This is an experimental indicator. The numbers below the word cloud represent topic shares are percentages of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on KAKEN, accessed August 2020.

NWO (NLD) funding

Table C.10 shows that the number of projects identified as AI-related nearly quadrupled from 30 in 2016 to 117 in 2019. The amount of R&D funding increased sixteen-fold from USD 5 million in 2016 to USD 80 million in 2019. The percentage of AI-related projects increased from 1.8% to 6.5%, while the percentage of funding allocated to AI-related projects increased from 1.5% to 12.3%, as shown in **Figure C.19**.

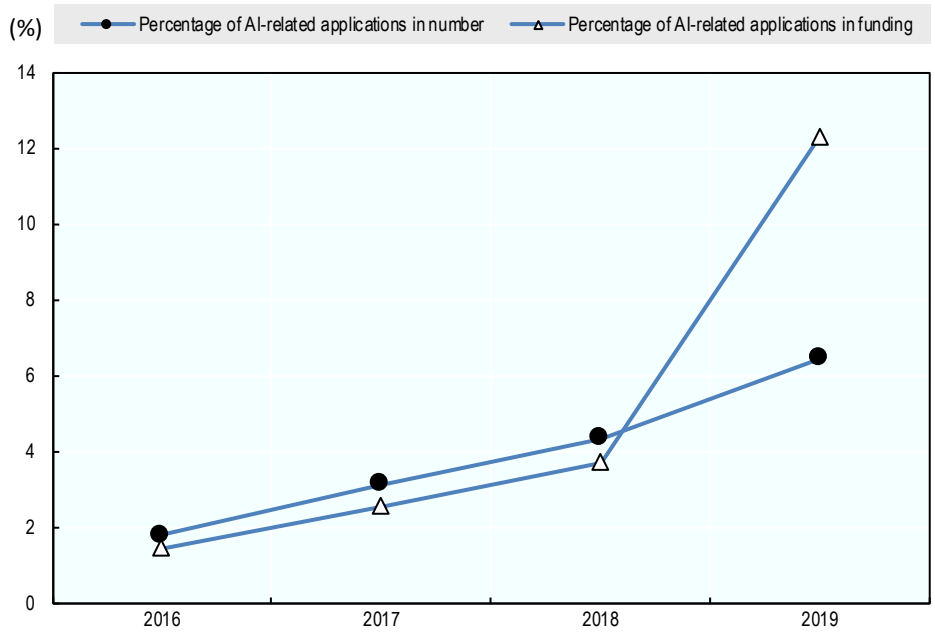
Table C.10. Estimates of AI-related R&D in NWO funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2016	30	1670	1.80	5	364	1.46
2017	59	1876	3.14	13	518	2.56
2018	79	1819	4.34	25	659	3.73
2019	117	1812	6.46	80	646	12.31

Note: These figures are results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: Rathenau Institute calculations by the codes developed by the OECD based on NWO data, accessed March 2020.

Figure C.19. Estimates of AI-related NWO funding

As a percentage of total NWO funding (number of projects and funding amounts)



Source: Ratenau Institute calculations by the codes developed by the OECD based on NWO data, accessed March 2020.

Figure C.20 shows the topics identified from the documents associated with the AI-related R&D projects funded by NWO. The topics retrieved were 1. Natural language processing, 2. Machine learning algorithm(s), 3. Dynamic power device(s), 4. Machine learning system(s), 5. AI technology in society, 6. Image analysis for treatment, 7. Software development, 8. AI robot development, and 9. Deep learning on brain.

Figure C.20. Topics from the AI-related documents of NWO with relative topic prominence, 2016-2019

Note: This is an experimental indicator. The numbers below the word cloud represent topic shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: Ratenau Institute calculations by the codes developed by the OECD based on NWO data, accessed March 2020.

NIH (USA) funding

Table C.11 shows the number of projects identified as AI-related has increased by ninefold from 2001 to 2019. The amount of R&D funding has increased more than fifteenfold in the same period from USD 53 million to USD 829 million. The percentage of projects identified as AI-related has increased elevenfold over this period, growing from 0.2% in 2001 to 2.2% in 2019. The percentage of NIH's total funding allocation given to AI-related projects displays a similar trend, increasing from 0.3% in 2001 to nearly 2.3% in 2019.

Figure C.21 shows a sustained growth in the share of AI-related R&D funding punctuated by a large singular increase in 2009, coinciding with the allocation of additional funds under the American Recovery and Reinvestment Act (ARRA). The comparison of project counts and funding data suggests that some of the projects classified as AI-related were relatively large. This is confirmed by the fact that there were exceptionally large applications submitted from 2009 onwards. As a result, applications with more than USD 100 million in funding were manually examined, as they have a large impact on the estimated share of allocated funding.

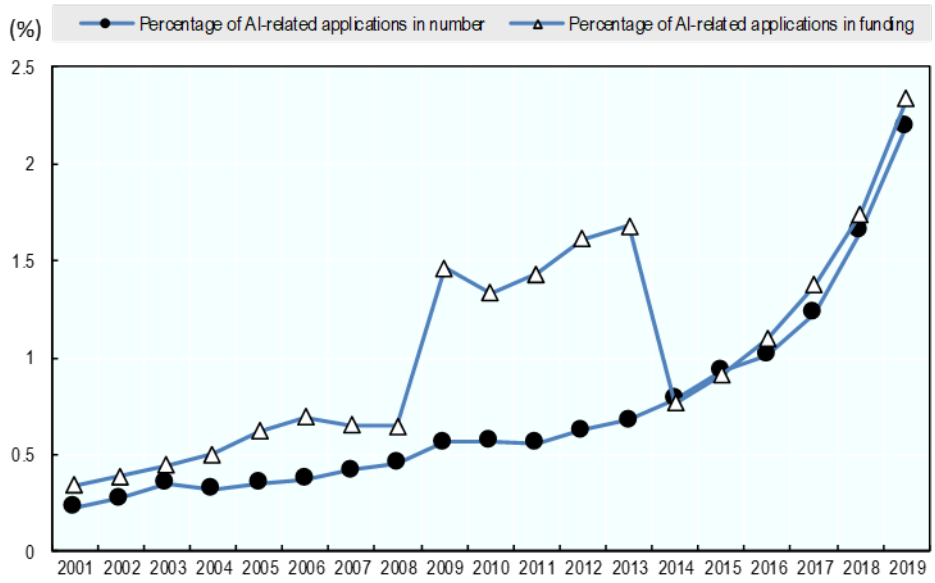
Table C.11. Estimates of AI-related R&D in NIH funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	180	79 248	0.23	53	15 665	0.34
2002	225	81 176	0.28	69	17 694	0.39
2003	209	59 174	0.35	88	19 789	0.44
2004	243	75 574	0.32	103	20 593	0.50
2005	276	78 227	0.35	133	21 249	0.62
2006	288	77 408	0.37	146	21 004	0.70
2007	348	82 940	0.42	173	26 443	0.65
2008	367	80 840	0.45	169	26 109	0.65
2009	516	91 088	0.57	465	31 798	1.46
2010	490	85 608	0.57	439	32 944	1.33
2011	418	74 585	0.56	398	27 755	1.43
2012	439	69 833	0.63	448	27 801	1.61
2013	459	67 679	0.68	442	26 313	1.68
2014	526	66 781	0.79	208	27 170	0.77
2015	626	67 043	0.93	249	27 257	0.91
2016	691	68 262	1.01	320	29 200	1.10
2017	861	69 956	1.23	422	30 747	1.37
2018	1 280	77 358	1.65	573	32 905	1.74
2019	1 661	75 692	2.19	829	35 519	2.34

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NIH RePORTER data, accessed August 2020.

Figure C.21. Estimates of AI-related NIH funding

As a percentage of total NIH funding (number of projects and funding amounts)



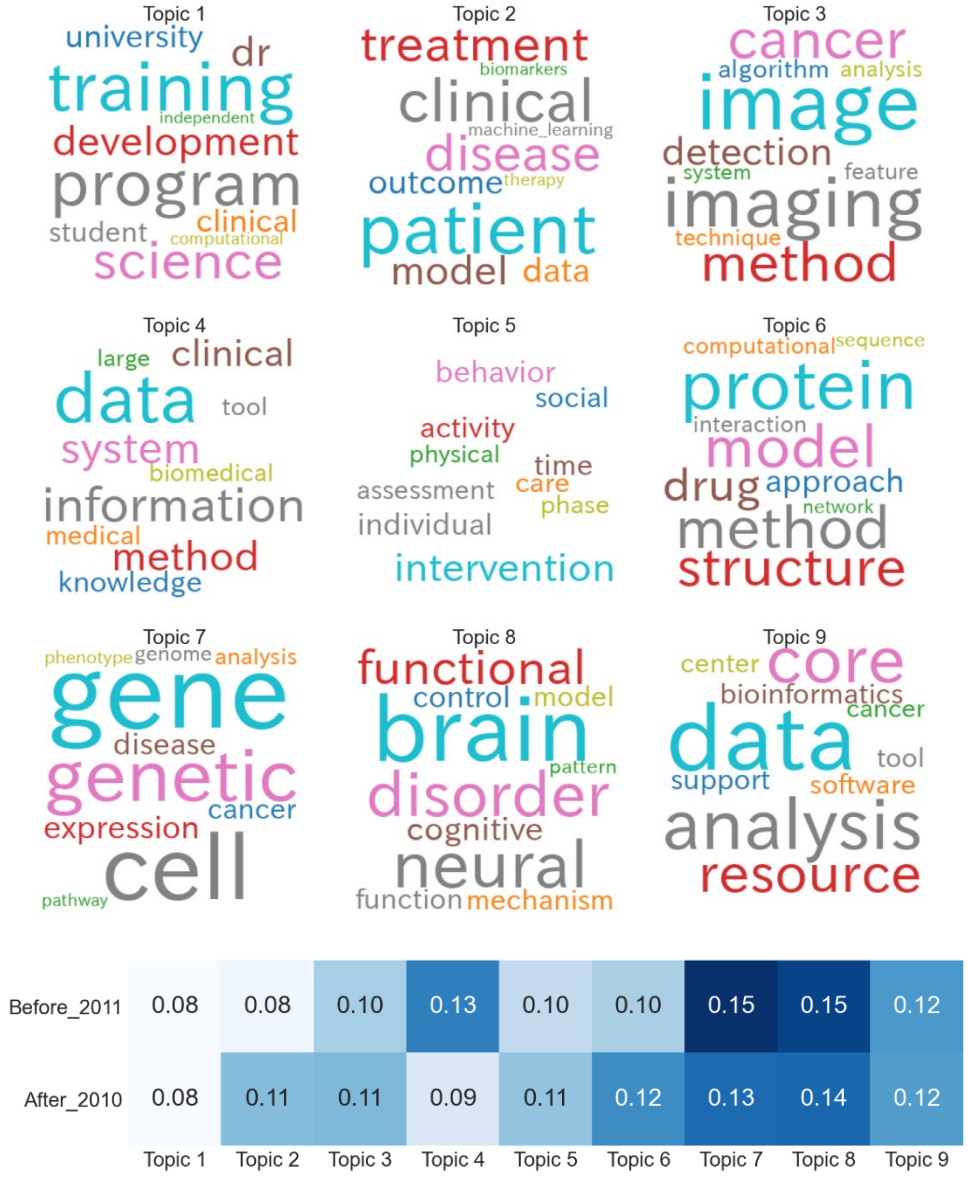
Source: calculations based on NIH RePORTER data, accessed August 2020

The extreme values in excess of USD 100 million are accounted for by an infrastructure project classified as AI-related. This project, called "National Biomedical Information Services", is an intramural project within one of the NIH institutes, the National Libraries of Medicine. This project has annual records with text descriptions that are very similar each year. It was decided that this project should be considered AI-related as it builds and provides various information systems that utilize AI systems, including natural language processing and tools for managing large bibliographic databases.

The results do nonetheless confirm that sustained growth in AI-related funding occurred over the period, despite a brief hiatus from 2006 to 2008. Growth was particularly rapid in the 2010s, in regard to both the quantity of projects and the quantity of funding allocated; there was no sign of deceleration in the final years for which data are available.

Figure C.22 shows the topics identified from the documents associated with the AI-related R&D projects funded by the NIH. The topics retrieved were 1. Training on data analysis, 2. Clinical application(s) of machine learning, 3. Image analysis for cancer detection, 4. Clinical data system(s), 5. Personalised care, 6. Protein structure analysis 7. Genetic disease analysis, 8. Brain function and disorder(s), and 9. Bioinformatics.

Figure C.22. Topics from the AI-related documents of NIH with relative topic prominence, 2001-2019



Note: This is an experimental indicator. The numbers below the word cloud represent shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on NIH RePORTER data, accessed August 2020.

NSF (USA) funding

Table C.12 summarises AI-related R&D funding by NSF. The number of projects identified as AI-related increased more than sixteen-fold from 97 in 2001 to 1 612 in 2019. The amount of R&D funding also increased from USD 47 million in 2001 to USD 721 million in 2019, a fifteenfold increase. Between 2001 and 2019, the share of AI-related funded R&D projects increased from 1.5% to 13.7%. A similar growth in the percentage of total allocated funding was observed: 1.1% to 14.5%. Both percentages surged after 2015; **Figure C.23** illustrates these trends.

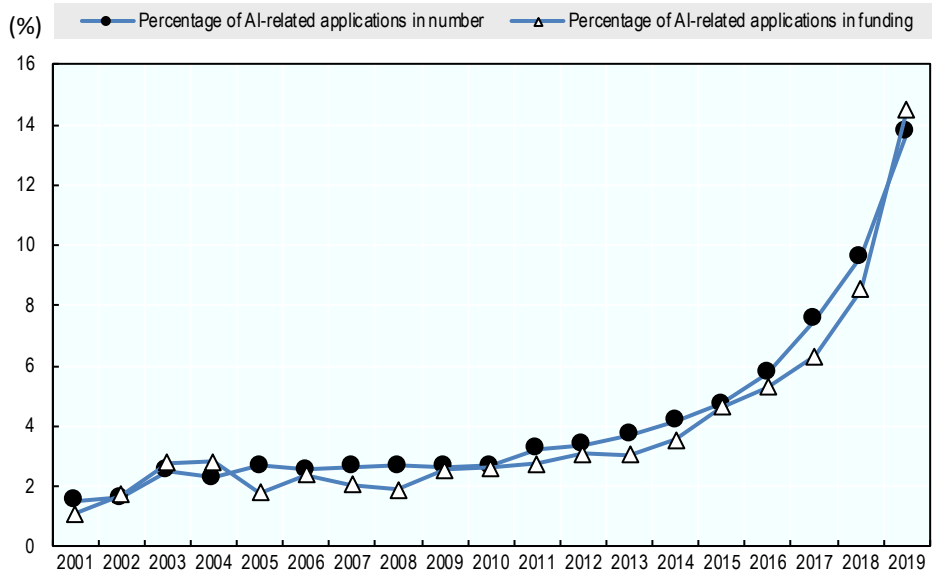
Table C.12. Estimates of AI-related R&D in NSF funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	97	6 330	1.53	47	4 200	1.12
2002	179	10 904	1.64	92	5 276	1.74
2003	293	11 584	2.53	181	6 457	2.80
2004	251	10 922	2.30	140	4 962	2.83
2005	280	10 405	2.69	92	5 118	1.80
2006	279	10 883	2.56	131	5 480	2.39
2007	320	12 115	2.64	104	5 027	2.08
2008	315	11 735	2.68	138	7 204	1.91
2009	403	15 233	2.65	206	8 029	2.57
2010	371	13 818	2.68	196	7 457	2.62
2011	387	11 947	3.24	183	6 672	2.75
2012	418	12 305	3.40	191	6 204	3.08
2013	434	11 701	3.71	176	5 786	3.04
2014	501	11 983	4.18	225	6 340	3.55
2015	612	12 926	4.73	268	5 750	4.65
2016	747	12 993	5.75	363	6 845	5.31
2017	921	12 197	7.55	408	6 468	6.31
2018	1 211	12 596	9.61	567	6 643	8.54
2019	1 612	11 730	13.74	721	4 964	14.52

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: OECD calculations based on NSF Award Search data, accessed August 2020.

Figure C.23. Estimates of AI-related NSF funding

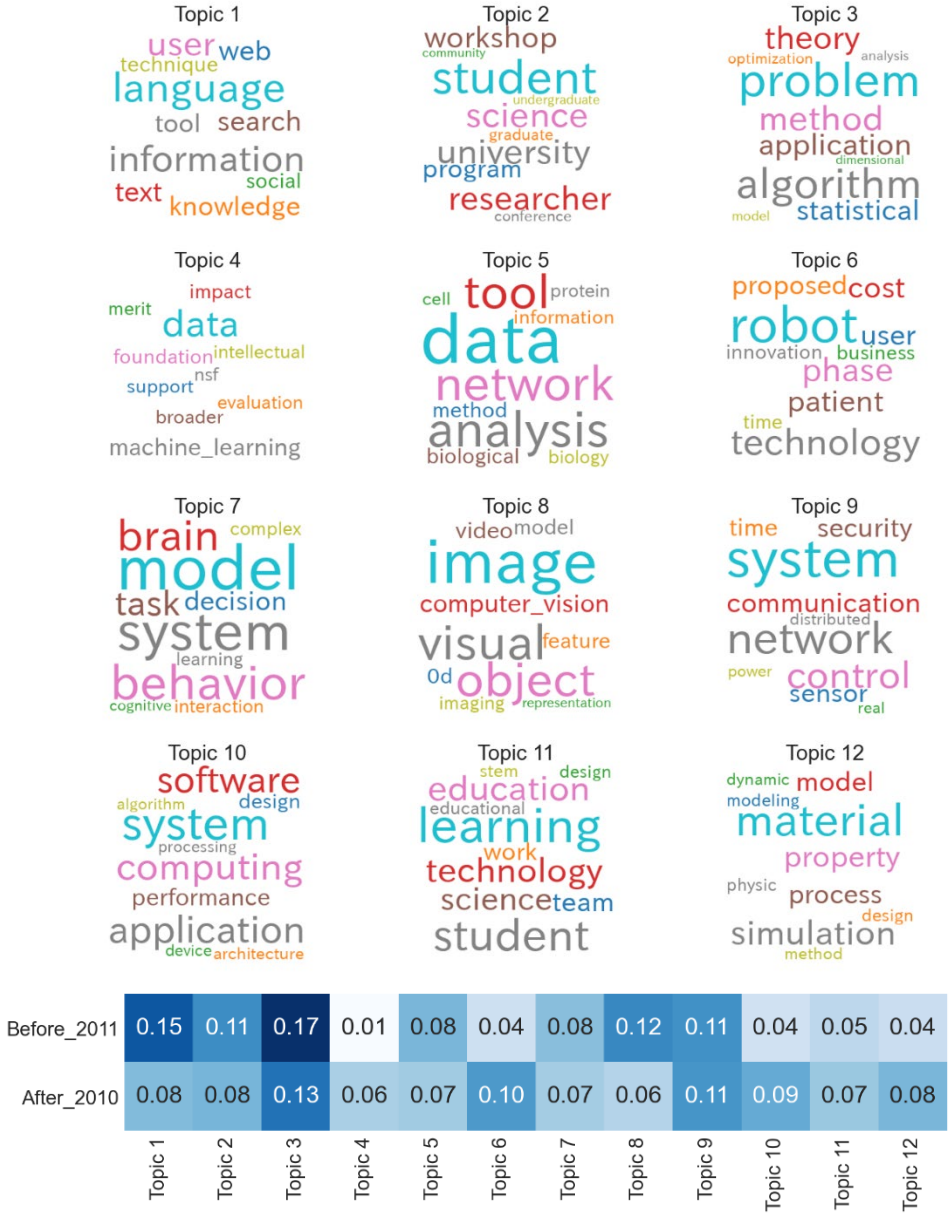
As percentage of total NSF funding (number of projects and funding amounts)



Source: OECD calculations based on NSF Award Search data, accessed August 2020.

Figure C.24 shows the topics identified from the documents associated with the AI-related R&D projects funded by the NSF. The topics retrieved were 1. Natural language processing, 2. Student training, 3. Statistical algorithms, 4. Machine learning application(s), 5. Network analysis in biology, 6. Robots for care, 7. Behaviour modelling, 8. Computer vision, 9. Sensor network systems, 10. Software and system development, 11. Technology education, and 12. Material analysis.

Figure C.24. Topics from the AI-related documents of NSF with relative topic prominence, 2001-2019



Note: This is an experimental indicator. The numbers represent shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: OECD calculations based on NSF Award Search (database), accessed August 2020.

CORDIS (EU) funding

Table C.13 shows that the number of projects identified as AI-related was stable up until 2014 before increasing nearly eightfold from 61 in 2014 to 472 in 2019. The amount of R&D funding gradually increased until 2017, fluctuating from USD 24 to 299 million, with a spike in 2005. In 2018, funding jumped to USD 649 million, followed by USD 1 163 million in 2019. The percentage of AI-related projects increased from around 1.0% to 8.5%, while the percentage of funding allocated to AI-related projects increased from around 1.0% to 9.3% as shown in **Figure C.25**.

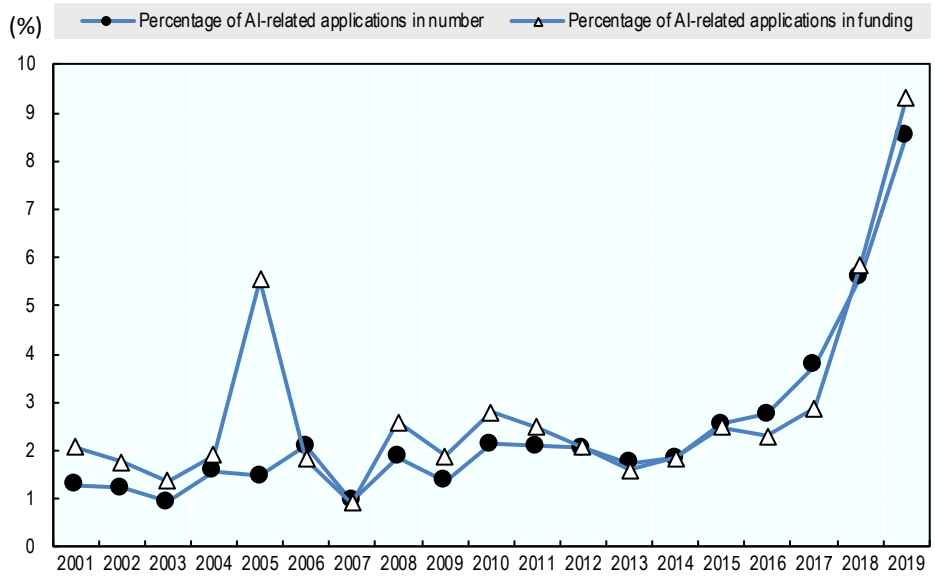
Table C.13. Estimates of AI-related R&D in CORDIS funding

	Number of granted applications			Funding amounts		
Year	AI-related projects	All projects	Percentage of AI-related projects (%)	AI-related projects (USDm)	All projects (USDm)	Percentage of AI-related funding (%)
2001	63	4 917	1.28	88	4 222	2.08
2002	55	4 461	1.23	80	4 533	1.77
2003	14	1 505	0.93	25	1 808	1.37
2004	40	2 527	1.58	124	6 495	1.91
2005	42	2 829	1.48	291	5 245	5.54
2006	65	3 089	2.10	137	7 530	1.82
2007	15	1 549	0.97	24	2 580	0.91
2008	54	2 903	1.86	173	6 712	2.58
2009	40	2 917	1.37	112	5 919	1.90
2010	77	3 606	2.14	207	7 391	2.80
2011	83	3 966	2.09	205	8 225	2.49
2012	89	4 332	2.05	192	9 209	2.08
2013	80	4 587	1.74	172	10 703	1.60
2014	61	3 303	1.85	119	6 507	1.84
2015	128	5 038	2.54	276	11 134	2.48
2016	137	4 955	2.76	245	10 638	2.30
2017	187	4 973	3.76	299	10 437	2.87
2018	285	5 080	5.61	649	11 100	5.85
2019	472	5 524	8.54	1 163	12 479	9.32

Note: These figures are the results of key term matching. In the case of multi-year projects, the number of applications and their funding amounts are assigned to the first year of operation.
Source: FECYT calculations by the codes developed by the OECD based on CORDIS data, accessed April 2020.

Figure C.25. Estimates of AI-related CORDIS funding

As percentage of total CORDIS funding (number of projects and funding amounts)



Source: FECYT calculations by the codes developed by the OECD based on CORDIS data, accessed April 2020.

Figure C.26 shows the topics identified from the documents associated with the AI-related R&D projects funded by the CORDIS. The topics retrieved were 1. Production technology, 2. AI application(s) to business, 3. Natural language processing, 4. Researcher training, 5. Sensor network(s), 6. Medical research, 7. Energy system(s), 8. Social change modelling, 9. Machine learning analysis, 10. Low power processing, 11. Algorithm development, and 12. Cognitive robot(s).

Figure C.26. Topics from the AI-related documents of CORDIS with relative topic prominence, 2001-2019

Note: This is an experimental indicator. The numbers below the word cloud represent shares of all AI-related projects in a specified period (e.g. if ten projects were classified as AI-related before 2011, three of which are associated with topic 1, the figure for topic 1 before 2011 would read 0.3). For each document, the topic with the highest probability of being associated to that document (as computed by the LDA algorithm) was selected.
Source: FECYT calculations by the codes developed by the OECD based on CORDIS data, accessed April 2020.

Annex D. Bias analysis and robustness checks

Analysis of potential bias

Incidence of false positives (false discovery rate)

In order to assess the robustness of the results, especially in light of the presence of somewhat ambiguous key terms that may bias these estimates (e.g. if two ambiguous AI terms or more are found in the same document), batches of 100 documents were selected at random from each of the NIH (10 103 projects) and NSF (9 631 projects) databases; these two databases served as representatives for all other databases containing life science R&D projects and general R&D projects, respectively. The titles and abstracts of these 200 documents were manually inspected and classified into four categories shown in **Table D.1**. Documents making clear reference to the use or development of AI systems (category A) can be treated as unambiguous true positives. In addition to this class, there are also projects, which from their descriptions, can be deemed to represent AI-related R&D activity, based on the context (category B). These are likely true positives. The remainder can be liberally considered as the measure of the false discoveries, covering instances in which the description makes it apparent that the project has little to do with AI (category D), as well as instance where not enough context is available to judge to what the selected Key AI terms actually refer (category C). Examples of projects allocated to different categories are provided in **Box D.1**.

Table D.1. Precision analysis of AI detection results in NIH and NSF data

Sample of randomly chosen documents from selected documents identified as AI-related

Status	NIH	NSF
A. Explicit AI relevance.	61	57
B. Likely AI relevance	29	37
C. Possible false positive (insufficient information in text to tell)	***8***	***4***
D. Likely false positive	**2**	**2**
Total	100	100

Source: OECD analysis based on NIH RePORTER and NSF Award Search data, accessed December 2018.

In the case of the NIH funding data, the more liberal estimate of the false discovery rate (FDR) is about 10% (categories C and D), and its more conservative estimate is 2%. In the case of NSF data, the liberal estimate of the FDR is lower, at 6%, possibly reflecting the fact that descriptions of project methodologies are more prevalent in NSF project abstracts than in NIH project abstracts. This discrepancy may also be due to as a prevalence of analogue terms like “neural network” being frequently used in neurological or developmental contexts (as opposed to AI contexts) in the NIH documents. Based on these samples, the conservative measure of FDR lies between close to zero and 7% for both the NIH and NSF with a 95% probability. The liberal measure lies in the 5-to-18% and 2-to-13% ranges, respectively.

This precision analysis allowed us to identify two instances (one in each database) where lemmatisation mistakenly converted a term into one of the AI terms, namely when "deeper learning" was converted into the generic "deep learning" term.[31] Furthermore, an instance of the term "supervised learning" was found to refer to actual instructional activity, not to supervised machine learning. This suggests that this term may need to be partly penalised in order to reduce the risk of false positives.

Box D.1. Excerpts from sample of projects automatically retrieved as AI-related

Projects with explicit AI relevance:

- **Neural Networks for Estimating and Compensating the Nonlinear Characteristics of Nonstationary Complex Systems.** "The approach is to use Echo State Networks and Simultaneous Recurrent Neural Networks with super fast learning algorithms (biological inspired algorithms such as particle swarm optimization), and other computational intelligence algorithms, to accurately measure the distortion by monitoring only voltage and current without the need for added transducers. Such fast and powerful **neural networks** could also be used for closed loop control of the offending nonlinear devices to mitigate the distortion." (NSF)
- **Automated NMR Assignment and Protein Structure.** "New algorithms and computer systems will be developed for determining protein structure from only four NMR spectra. The system will use algorithms similar to and adapted from physical geometric algorithms, **pattern recognition** and **machine vision**, signal processing, and **robotics**, in order to analyze spectra, assign spectral peaks to atom interactions, compute secondary structure, and estimate the global fold." (NIH)

Projects that appear to be incorrectly or ambiguously identified as AI-related:

- **Identification and characterization Of A Complex Involved In Bloom Syndrome.** "In **bioinformatics** searches of the human genome, we noticed that human genome contains proteins with OB-fold domains similar to those in RMI and RPA." (NIH)
- **Recombinant Multiepitope Mosaic Protein Design for Urine-based Diagnosis of Leptospirosis**. "Our approach involves the use of computational biology and **bioinformatics** to create, score, and select "mosaic" antigens from Leptospira spp. Antigenic properties of the mosaic antigens are evaluated by indirect ELISA using a panel of well-characterized human sera from clinical patients and apparently healthy individuals. We will then use recombinant DNA and protein engineering techniques to derive cognate chimeric proteins." (NSF)
- **Personalized sensor based digital media simulations for Biology and Health education.** "In this project, we present a set of sensor-enabled, multimodal, NGSS aligned, validated formative assessment games for biology and health education. Our emphasis will be on high engagement, **deeper learning** of the heart and cardiovascular function and

[31] However, one of such instances may actually refer to an AI-related project because, as highlighted in one of the examples, types of sensor-enabled games may be supported by AI systems. Further search confirmed that the company that was awarded the funding reports on its website that it "transforms training, sales, service, production, and design by leveraging virtual and augmented reality (VR/AR), simulation, sensing, artificial intelligence, and machine learning (AI/ML) across the totality of employee, customer, and product life cycles."

diseases and valid formative feedback to guide next steps for teachers and to allow student to assess themselves". (NIH)

CAREER: **Compiler and Runtime Support for Multi-Tasking on Commodity GPUs.** "GPU computing has become mainstream, as witnessed in various domains such as **machine learning**, graph analytics, and scientific simulation. This CAREER project aims at developing a set of compiler and runtime techniques to support multi-tasking on commodity GPUs in a transparent and efficient manner" (NSF)

Incidence of false negatives (false omission rate)

To assess the potential error associated with using an incomplete list of key terms, a similar manual detection process is followed for documents not identified as AI-related. In order to determine what percentage of such documents were incorrectly excluded by the AI tagging procedure, 100 documents each were selected at random from both the 1 483 897 NIH documents and the 214 676 NSF documents categorised as non AI-related.

In the NIH case, none of the 100 projects examined were classified as type A or B (i.e. AI-related R&D). 10 documents were categorised as C (i.e. as possible false negatives) and 90 as D (i.e. as likely true negatives). Very similar results were obtained for the NSF sample. These results imply a low false omission rate, in the order of 5 to 18% on the most liberal measure, which is a likely overestimate (and 0 to 5% on a more conservative measurement) with a 95% confidence probability.

Table D.2. False omission analysis of AI detection results in NIH and NSF data

Categories	NIH	NSF
A. Explicit AI-relevance (false negative)	**0**	**0**
B. Likely AI relevance (likely false negative)	**0**	**1**
C. Possible false negative (insufficient information in text to tell)	**10**	**8**
D. Likely true negative	90	91
Total	100	100

Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed December 2018.

The rejected documents assigned to the C category after examination (i.e. the possible false negatives) reveal some of the challenges of determining AI relevance (or for that matter, the use or relevance of any other technology) when adoption rates are rapidly increasing. The example of a project untagged by the chosen key terms that seeks to "add strength in statistical methods for genetic data, clinical prediction, and paediatric oncology" raises questions as to whether, and under what circumstances, it is plausible for statistical methods not to be AI-enabled. One should bear in mind that adding selected complementary inputs (genetic data) and applications (clinical prediction) into the selection process risks overly extending the selection window.

General assessment

While it may appear that there is worse false discovery rate than false omission rate, it is important to note that the false discovery rate estimate applies to what is currently a

much larger group, so ultimately the results may actually underestimate the extent of true AI-related R&D funding by quite some margin.

After examining 400 documents to estimate the false discovery and omission rates, there are some potential lessons for refining the keyword matching approach. Firstly, the core set of AI-related terms appear to provide reasonably unambiguous predictors for AI relevance (e.g. "machine learning", "natural language processing", and "deep learning"). Even the problems associated with the use of the term "deep learning" in education science projects were found to be relatively minor, as such projects were often found to have been carried out by AI experts and/or to have resulted in AI publications. The documents that contain core AI terms could be extended provided that the text mining techniques used can cope with very large vocabularies.

A distinctive feature of key term identification in funding data, compared to working with counts of projects or documents, is the importance of prioritising the assessment of large projects, as overall funding totals might be skewed because of inaccurate measurements of individuals projects funded to the tune of tens of millions of USD or more. Larger projects also present difficult choices regarding whether the entirety of or only a fraction of the amount allocated should be treated as AI-related funding. These are questions that lie beyond our current scope, but which necessitate further consideration.

Furthermore, potential but non-exclusive terms used for AI should be further checked against the context in which they are used on a document-by-document basis, progressively substituting for the simple scoring approach. As noted, our approach requires some degree of human discretion in relation to key terms before applying the simple naïve rule. Results can also be rather sensitive to the decision rule of how many non-core terms to require.

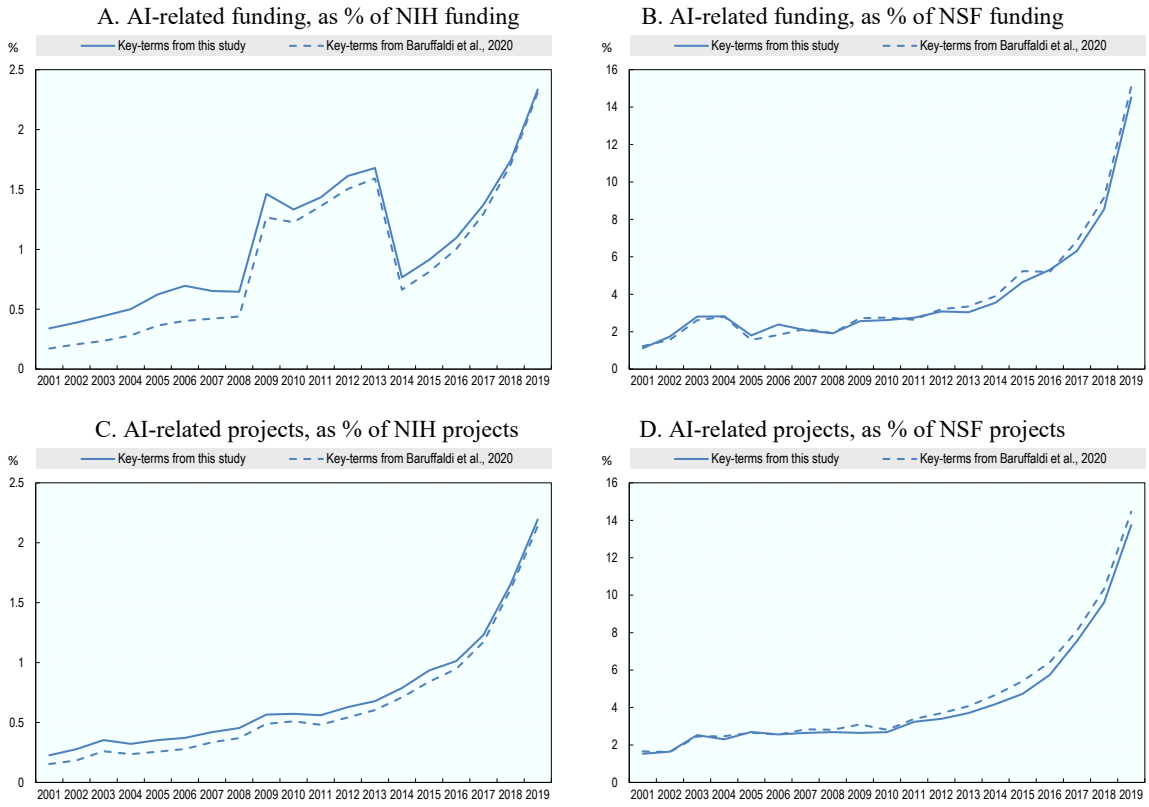
The robustness test based on an assessment of random samples of the corpus reveals the fundamental inconclusiveness of many project abstracts, particularly, but not exclusively, for the projects identified as non AI-related. This points to the fundamental challenge of the fitness of purpose of the abstracts. Abstracts are nonetheless convenient to use because they keep the data size manageable, but also because they are more accessible than detailed project descriptions, which are unlikely to qualify for public access.

Results are fairly robust to the use of alternative lists of key terms

As a further robustness test, these results have been compared to those that would be derived using an alternative (and longer) final list of key AI terms. The measurement of AI-related activity has been considered in recent OECD work (Baruffaldi et al., 2020[18]). The list of 193 terms derived in that context for the purpose of analysing the corpus of scientific publications (including conference proceedings) and also of being applicable in other contexts provided a useful check on the sensitivity of estimates to alternative methods that compile lists of identifying key terms.

A comparison of both lists of terms showed that while they overlap significantly (38 common terms), the Baruffaldi et al. (2020) report contains a significantly longer list of terms. Our study does however contain a smaller but still significant number of key terms that are not present in the comparison list. This suggests that it is possible to reduce further the risk of false negatives by combining both lists, although this might heighten the rate of false positives. "Bioinformatics" is the most common term in this paper's list that is not found in Baruffaldi et al. (2020), while "computational model" is the most common term in that report that was not included in this paper's list. Quantitatively, the application of the different lists results in similar estimates of AI-related project counts and funding amounts, as shown in **Figure D.1**. To make a like-for-like comparison, the same criterion of at least one core or two non-core terms was applied for both lists.

Figure D.1. Illustration of sensitivity of results to using alternative AI term lists



Note: The same decision criterion has been applied for the estimates applying to each list of key terms.
Source: OECD calculations based on NIH RePORTER and NSF Award Search data, accessed August 2020, and key AI terms in Baruffaldi et al. (2020).

In the case of NSF funding, estimates in this paper are on average slightly lower than those obtained when using the longer list. This gap is clearer for data collected after 2011. Between 2012, estimates are virtually identical. Conversely, for NIH funding, estimates based on this paper's list are higher before 2018, with the gap closing afterwards, and the figures for 2019 appearing nearly identical. This suggests that the health-oriented NIH corpus is quite distinct from general funding databases, and the specific semantic analysis of this corpus allowed us to retrieve more potentially relevant documents and funding amounts, especially prior to 2018. Indeed, project counts estimates are more similar between lists than funding amount estimates.