Activity Deliverable

Cross-KIC Artificial Intelligence

Report on assessment of data platform maturity and socially sustainable use of algorithms

EIT Manufacturing is supported by the EIT,
a body of the European Union

Reporting Year:	2020
Activity Code:	20185
Deliverable ID:	D01
Area Name:	7. Cross-KIC
Segment Name:	7.9 Cross-KIC Artificial Intelligence
EIT-level KPIs:	-
Contributing Partners:	000 EIT Manufacturing 000_2 CLC Central

Version History:

Version	Date	Owner	Author(s)	Changes to previous version
1.0	2021-01-27	Christian Bölling	Timo Scherer Christian Bölling	Initial submitted version

Contents

Executive Summary

Within the Innovation Impact AI cross-KIC project, EIT Manufacturing is responsible for work package 2, focussing on the topics data and algorithms, two aspects essential to creating powerful artificial intelligence (AI) applications. Data is an increasingly valuable good while at the same time becoming more widely available. Data management is therefore becoming more important, however, depending on the industry a business operates in, data management expertise is likely to be a rare commodity and data management itself might be more complex, especially when the business processes are rooted in the physical than in the digital world. Still, the efforts of managing data correctly and effectively deem a reward, since data fuel algorithms and therefore AI applications.

This delicate interplay between data and algorithm was the centre point of EIT Manufacturing's work package, with the goal to determine the maturity and level of advancement among members of the KICs' network. For this purpose, an online survey was created and shared among the KIC members, thereby distinguishing between members looking to adopt AI solutions (AI solution adopters), those providing businesses with AI solutions (AI solution providers), and members investing in AI technologies and related start-ups (AI investors). Those three domains of respondents were presented with (slightly) varied versions of the developed questionnaire, to assess their (adopters) or their customers' (providers) situation on and approach to various issues i. a.:

- Data governance body functionality (availability, accessibility, quality, and integrity of data)
- Data privacy and security measures
- Usage of (multiple) data platforms and related issues
- Used algorithms and application fields
- Protection against algorithm-based discrimination
- Barriers and drivers of AI integration

The survey ran over the summer month and collected answers from 60 KIC affiliates, with the domain of AI investors being underrepresented (2 responses) and therefore not meriting sound results. However, comparing answers from AI solution adopters and providers gives interesting insights to variations in situation assessment, which could lead to communication issues and therefore hamper progress of AI implementation projects.

1. Data and algorithms – foundation of AI applications

AI and especially machine learning applications are mainly dependent on two things: data and algorithms. Sufficient amounts of data, that are available in the required quality, accessible when needed, and trustworthy is the input required to train and run machine learning algorithms. The quality of AI-based decisions and therefore the impact of AI applications is strongly dependent on data, which is why subjects like data governance and management are becoming increasingly important to businesses attempting to incorporate AI into their workflows. At the same time, businesses with less digital backgrounds (e. g. manufacturing, health) often struggle dealing with data, AI, and digitalisation topics overall. They therefore rely on external services to manage and handle their data (platform solutions) and approach experts for help to develop or customize machine learning algorithms to integrate AI into their workflow.

Reliance on external resources is a valid approach, given that the expertise necessary to implement digitalisation measures and AI learning applications is a rare commodity. However, handling of data platform solutions and interaction with solution providers brings challenges in itself. Due to differences in technical background and expertise, communication between adopters and providers of AI solutions often do not occur eye-to-eye, increasing the number of iterations required to finish certain tasks, hence increasing project runtime, cost, and also frustration among parties. Determining the areas in which misconceptions and misunderstandings occur and to what degree offers an opportunity to establish measures to mitigate these issues and contribute to accelerated AI adoption by businesses.

Besides businesses looking to adopt digitalisation and AI solutions (AI solution adopters) and those providing their services (AI solution providers), a third domain of players has substantial interest in ensuring the success of AI technologies and related businesses – AI investors. To analyse the viewpoints on data, algorithms, and related issues of these three domains in order to determine discrepancies, an online survey was developed.

2. Online survey among different domains

As described in section 1, an online survey was designed to assess the viewpoints on data, algorithms, and related issues among the following three domains of respondents:

- AI technology adopters
  organisations that (plan to) use AI to improve their processes

- **AI solution providers**
  organisations offering AI solutions and applications
- **AI technology investors**
  organisations that invest in AI and/or AI solution providers

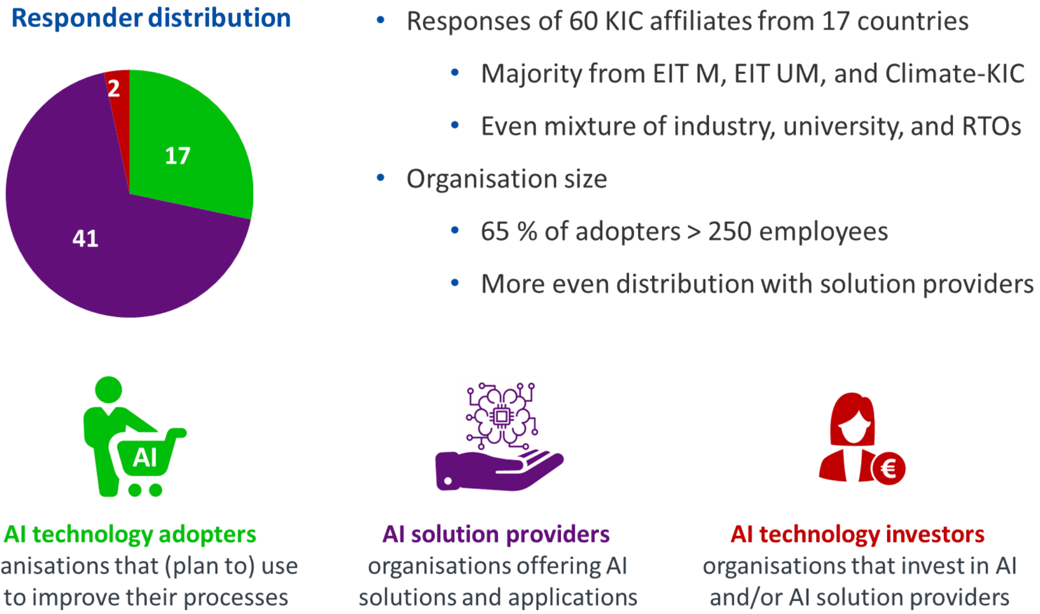
In close collaboration with EIT Urban Mobility, a questionnaire with overall 61 questions was designed, with respondents from each domain receiving variations of questions, either focussing on their in-house situation (adopters) or the situation with their customers (providers). This distribution of focus points between the different domains led to an online questionnaire, that was hosted via the SurveyMonkey online platform, that took between 5 and 9 minutes to complete.

The questionnaire was structured as follows:

- A section asking for background information on respondents to learn about the demographics participating in the survey
  - Among other things, respondents were asked about their position, KIC affiliation, years of professional experience, field of study, type of organisation, country the organisation is located in, number of employees, etc.
- A section asking about the data governance situation in the respondent's organisation or with the respondent's customers
  - Overall rating of data governance body functionality, degree of data availability, data quality, data integrity, data security
- A section dealing with data privacy topics such as compliance with regulations, risk-based data categorisation, commitment of management towards data privacy topics.
- A section asking about the role data platforms play in organisations, which platforms are being used, simultaneous usage of multiple platform solutions and problems occurring from this.
- A section about algorithms and their application field, as well as the problem of potentially discriminating algorithms.
- Additional sections of the survey dealt with AI solutions, their impact, barriers, and catalysts, as well as business models that originate from using AI. (This section was mainly part of work package 3 (EIT Urban Mobility) and was integrated in the survey to create synergy effects).

The survey design was mainly of select or multiple-choice character, with few exceptions asking respondents to answer freely. To disseminate the survey and attract respondents, various communication channels of the involved KICs were used (mass email to partner master contacts, newsletters, social media postings, etc.). The survey was online for multiple weeks and collected 60 responses (the survey was published over summer vacation time, otherwise an even higher number of responses would have been possible).

The majority of respondents answered from the AI solution provider viewpoint (41), whereas 17 respondents answered from an AI technology adopter perspective. Unfortunately, only 2 AI technology investor answers were received, hence not yielding truly significant results. Majority of respondents are affiliated with EIT Climate KIC, EIT Manufacturing, and EIT Urban Mobility, but answers from EIT Digital, EIT Health, EIT Inno Energy, and EIT RawMaterials were also collected. Overall, responses from 17 countries were received, with the majority coming from Germany, Spain, the United Kingdom and France. Figure 1 summarises the demographics of respondents.



Responder distribution

- Responses of 60 KIC affiliates from 17 countries
  - Majority from EIT M, EIT UM, and Climate-KIC
  - Even mixture of industry, university, and RTOs
- Organisation size
  - 65 % of adopters > 250 employees
  - More even distribution with solution providers

AI technology adopters
organisations that (plan to) use AI to improve their processes

AI solution providers
organisations offering AI solutions and applications

AI technology investors
organisations that invest in AI and/or AI solution providers

Figure 1: Summary of demographics of survey respondents

3. Summary of survey results

3.1 Data governance situation

In the questionnaire, data governance (DG) was defined as the overall management of the *availability*, *usability*, *integrity* and *security* of the data you use. Following this definition, adopters and providers were asked to evaluate the perceived condition of the DG within one's organisation, or within their customers' organisation, respectively. Results show that AI technology adopters tend to rate their DG bodies' condition far more positively than providers of AI solutions. Whereas more than half of adopters would describe their DG body as functional albeit featuring room for improvement, nearly half of solution providers state that most or rarely any of their customers feature a functional DG body.

To gain deeper insights to the data governance situation, respondents were asked to provide details on data availability, which describes to what degree data is available to end-users and applications when and where they need it. Asked to categorise data availability within their organisation, or their customers' organisations, respectively, from poor to excellent, AI technology adopters and solution providers disagreed continuously:

- Whereas none of the adopters would describe their data availability as poor, nearly a fifth of solution providers would describe the situation within their customers that way.
- Assessments of medium or limited data availability diverge less among both domains, with little over half of solution providers describing their customer situation that way, compared to 38.5 % of adopters.
- The largest discrepancy is found in the "good data availability" category, described as "most data being available most of the time". The clear majority of 54 % of solution adopters would agree with this statement compared to only roughly a fifth of solution providers.
- Finally, 6 % of solution providers describe their customers' data availability as excellent, with any data being available at any time – an assessment shared by none of the technology adopters.

The results are also depicted in the Figure 2.

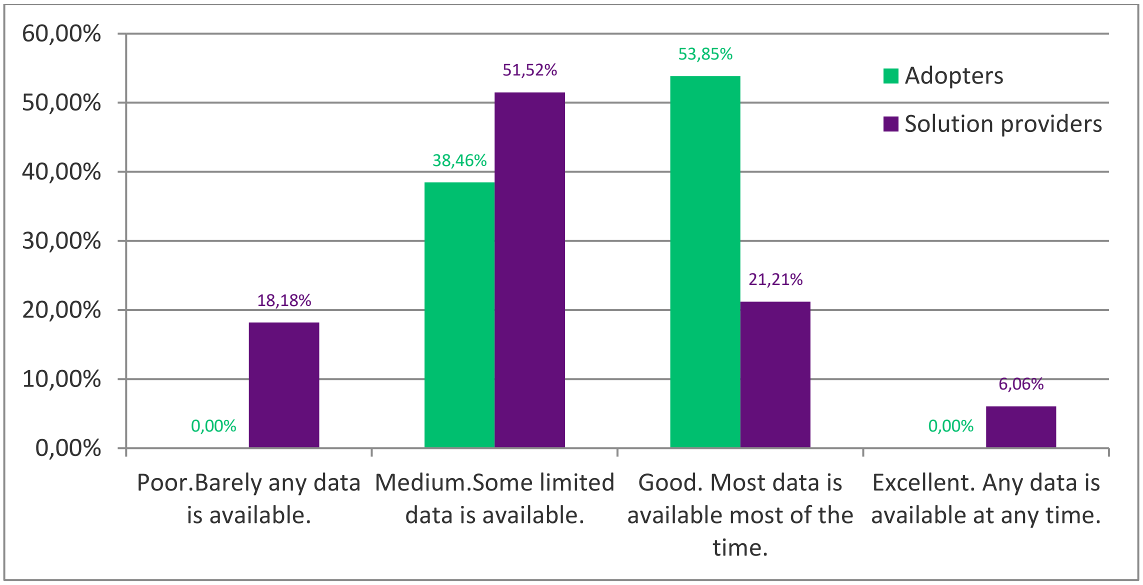


Figure 2: Assessment of data availability

Regarding data integrity – the accuracy, consistency, and validity of data over its lifecycle – assessments between both domains are quite similar. Respondents were asked to rate the situation within their or their customers' organisation from "poor" to "excellent":

- 0 % of technology adopter describe their data integrity as poor, compared to more than a tenth of solution providers with regard to their customers' data integrity.
- Describing data integrity as medium – varying strongly and requiring checking before usage – has the broadest consensus among domains, with 46 % of adopters and 45 % of providers, assessing their or their customers' data integrity in that manner.
- Respondents describing data integrity as "good" however shows slight discrepancy between domain, with 46 % of adopters and only 36 % of providers sharing the assessment that most of the data is reliable.

- Data integrity rating of “excellent” is rare among both domains, with 0 % of adopters and 3 % of providers assessing that all data is reliable.

To evaluate an organisation’s data security condition, standards, and technologies in place to protect data from intentional or accidental destruction, modification or disclosure need to be considered, along with technologies that limit access to data to unauthorised or malicious users/processes. Assessment of data security situations within their or their customers’ organisations were in part highly disputed between technology adopters and providers:

- Rarely any response from either domain indicates poor data security situations, with 0 % of adopters and only 6 % of providers stating that data breaches are occurring on a regular bases within one’s or customers’ organisation.
- The largest discrepancy between domains exist in attesting the data security situation to be medium, describing an improving situation with security problems in the past. Again, 0 % of technology adopters would grade their data security situation as “medium”, compared to almost half of technology adopters rating their customers’ situation in that manner.
- Nearly 70 % of adopters describe their data security situation as good, stating that data security has never been an issue, an assessment shared by merely 30 % of technology providers.
- Finally, describing data security standards in their or their customers’ organisation as excellent (striving to always implement latest data security technology) reveals less discrepancy between the domains: 23 % of adopters vs. 12 % of providers.

The survey results regarding data security are depicted in Figure 3.

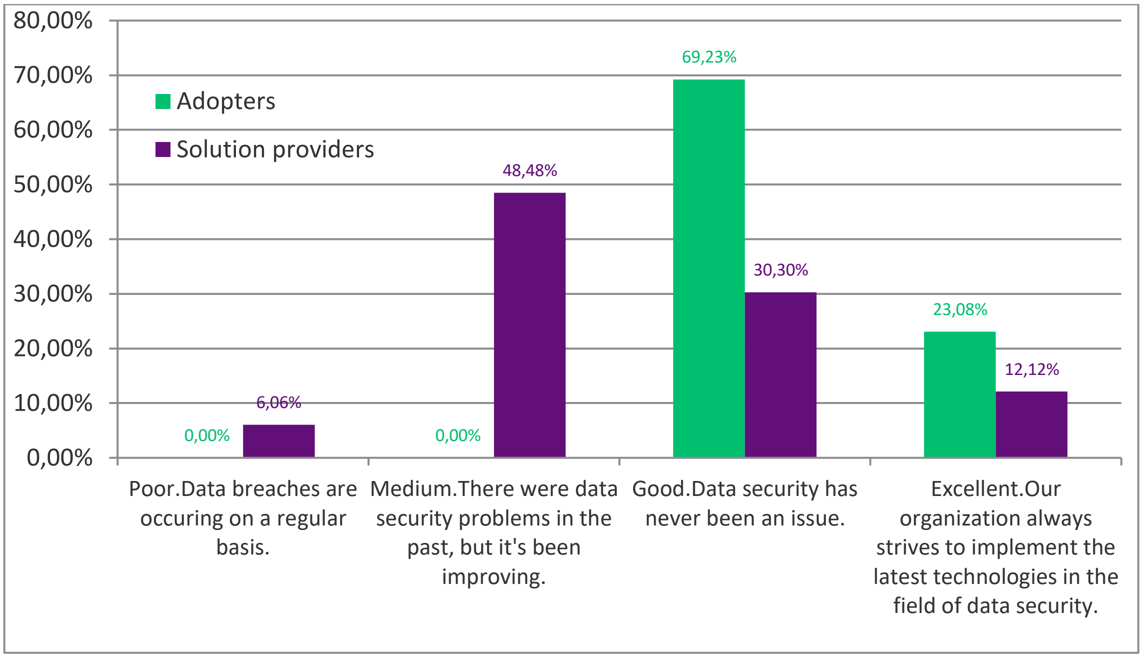


Figure 3: Assessment of data security

Strongly related to data security is data privacy, with the distinction that privacy issues are primarily of legal nature whereas security issues are of technical nature. Data privacy describes practices which ensure that data shared by customers is only used for its intended purpose. Privacy therefore deals with data usage given authorised access, compared to security dealing with preventing unauthorised access. Data privacy is a critical topic and therefore subject to regulations on states, national, and international level. Respondents of both domains were asked to describe the degree to which these regulations are followed. (Note: in contrast to previous data governance questions, providers of technology solutions were asked to rate their own handling of data privacy matters, not their customers'.) Both domains respond comparably, with 0 % (6 %) of technology adopters (providers) state to not follow any regulations, and 15 % (6 %) of technology adopters (providers) state to follow them at least partially. The vast majority of both domains is following data privacy regulations (adopters: 77 % | providers: 79 %).

Complying with data privacy regulations requires certain efforts, which should be backed by organisations' management bodies. Respondents of the technology adopter domain were therefore asked to rate their management bodies' commitment to ensure data privacy is given. Responses show that data privacy is an important topic for managers of technology adopters, with 23 % stating that management shows at least "fair" commitment to the cause. The majority of 70 % of responses stated that their management shows full commitment to data privacy issues.

Furthermore, to structure data privacy issues, risk-level classification of data is a common approach and mandatory depending on the type of data. The following three risk-levels are distinguished:

- Low risk: data is intended for public disclosure
- Moderate risk: data not generally available to the public
- High risk: data protection is required by law

When asked if they performed data categorisation based on risk levels, one third of responding technology solutions stated to do so partially, whereas two-thirds use risk-level data categorisation throughout their organisations' data.

3.2 Data platforms

Data platforms exist in manifold ways, featuring vastly different functionalities, and are used for different applications. One possible definition of data platform is an integrated technology solution that allows data to be governed, accessed, and delivered to users, data applications, or other technologies for strategic business purposes. Data platforms are used by technology solution adopters and providers alike, the latter one might even use data platforms internally and at the same time as part of their service offerings. Given these different use-case scenarios, the survey looked to determine differences in requested data platform functionality between adopts and providers of AI technology solutions. Asked to rate different functions of data platforms from slightly important to very important, both domains proofed to have comparable needs.

Four of the five top rated data platform functions were similar between solution adopters and providers. Data platform usability was the top priority for adopters, probably since handling and operating platforms can be complex and, as described above, adopters may lack staff with the required expertise. Usability was ranked third among solution providers. Top priority for solution providers, and runner-up for adopters, was data platform security, with 81 % (providers) and 91 % (adopters) ranking it (very) important. Cost for data platform operation and analysis/reporting capabilities are other top functionalities valued by both domains. Discrepancy regarding the top five data platform functions exist with regard to process automation and data visualisation, with the former being a top priority for adopters and the latter for solution providers. The summarised results for preferred data platform functionalities is depicted in Figure 4.

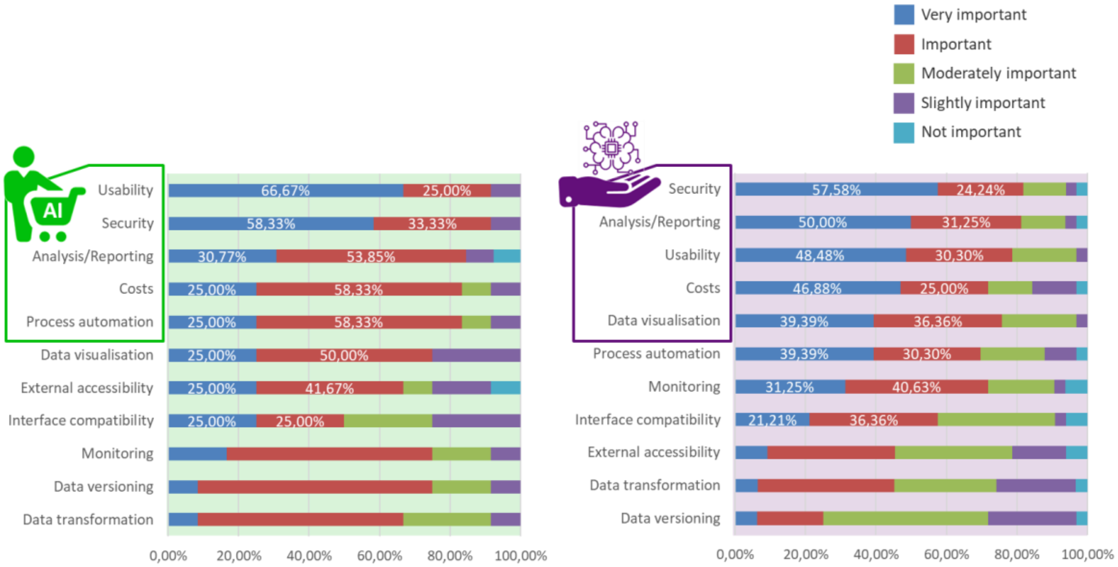


Figure 4: Rating of data platform functionality among adopters and providers of AI technology solutions

There is a large variety of data platform operators and service providers, ranging from giant tech companies like Microsoft and Google, to smaller and lesser-known operators like for example Datameer. To describe to what degree technology solution adopters are interacting with certain data platform operators, respondents from that domain were given a selection of more than 10 established data platform operators. Results show that adopters had at least heard about all platforms mentioned. Especially the Dell Big Data Analytics platform is well known with over 83 % of adopters having heard of it. However, despite having been in contact with over 16 % of responding adopters, none of them decided to work with Dell as platform operator. The platform solution from Intel shows comparable characteristics. Platform solutions from Microsoft (36 %) and Google (33 %) are used most often by technology adopters, followed by Amazon and Tensorflow with 25 % each. Given that the majority of responses were made by affiliations of the EIT Manufacturing KIC, the fact that none of the adopters stated to work with the Siemens Mindsphere platform solution, is a bit surprising. The distribution of degree of interaction of AI technology adopters with data platform solutions is summarised in Figure 5.

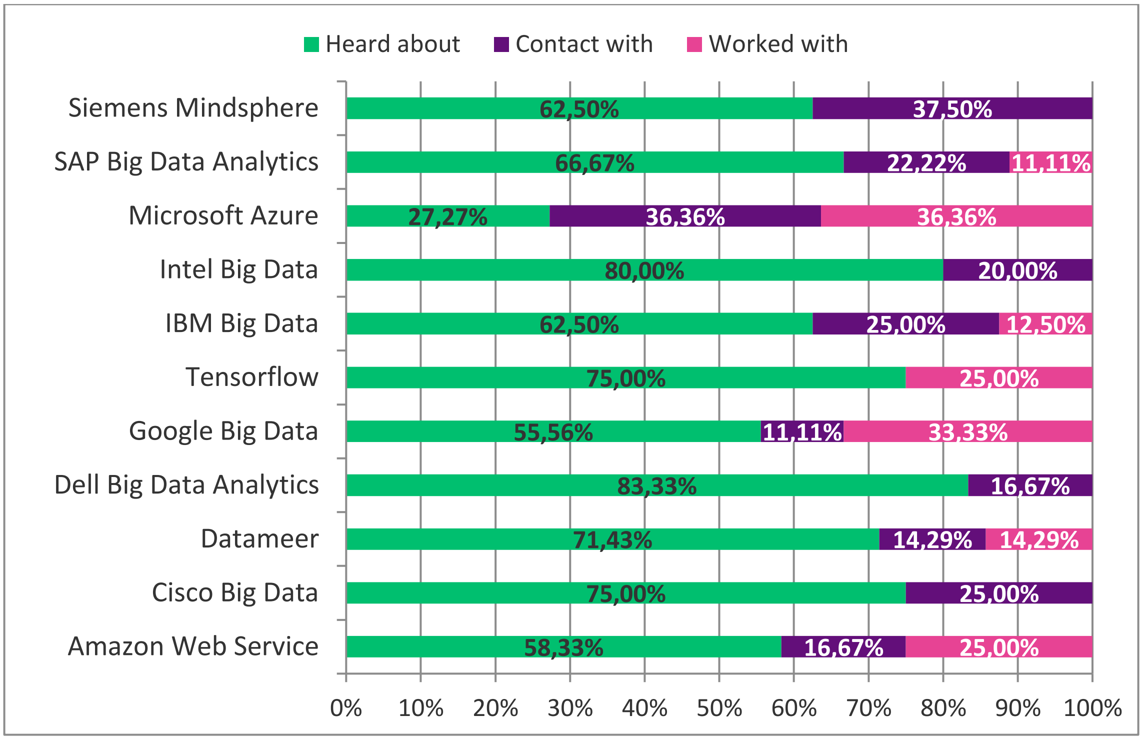


Figure 5: Degree of interaction with data platform operators (AI technology adopters)

Given the variety of data platform operators and functionalities, questions if adopters of technology solutions rely on a single or on multiple data platforms, and if they encounter interoperability issues arise. Asked about this, none of the responding AI technology adopters stated to not use any sort of data platform, and only 15 % stated to only use one individual solution. The vast majority (over 60 %) states, that they are indeed using multiple platform solutions and are in fact encountering issues due to lacking interoperability. However, none of the respondents state that they are therefore planning on switching to single-platform operation, showing that companies are willing to live with the challenges and issues, indicating that the value proposed by using multiple platforms outweighs the troubles. Finally, a small share of less than 8 % of technology adopters state to use multiple platforms and to never have had problems at all. The distribution of experience dealing with data platforms is summarised in Figure 6.

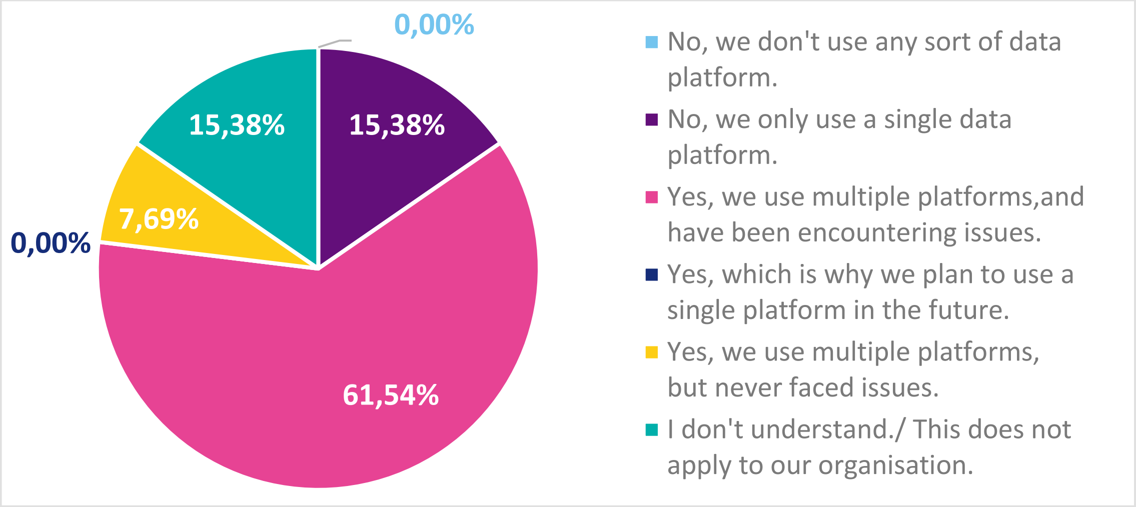


Figure 6: Experience of AI technology adopters on data platform usage

3.3 AI / Machine Learning Algorithms

Among other things, providers of AI technology solutions offer the development, adaptation, and implementation of algorithm-based solutions to their customers. Depending on among other things available data, intended application field, and expertise available, solution providers rely on a selection of machine learning algorithms. In order to determine which algorithms are used most frequently, solution providers were provided with a selection of 12 common algorithms to choose from. The results show two clear favourites among solution providers, with neural networks (86 %) and decision trees (76 %) being used by a clear majority of responding technology providers. Third most common are so called evolutionary and genetic algorithms that are used by little over half of solution providers. The survey also shows that practically all responding providers of AI technology solutions in fact rely on algorithms, with only a little over 3 % stating to use algorithms at all. Figure 7 summarises the results of technology provider algorithm usage.

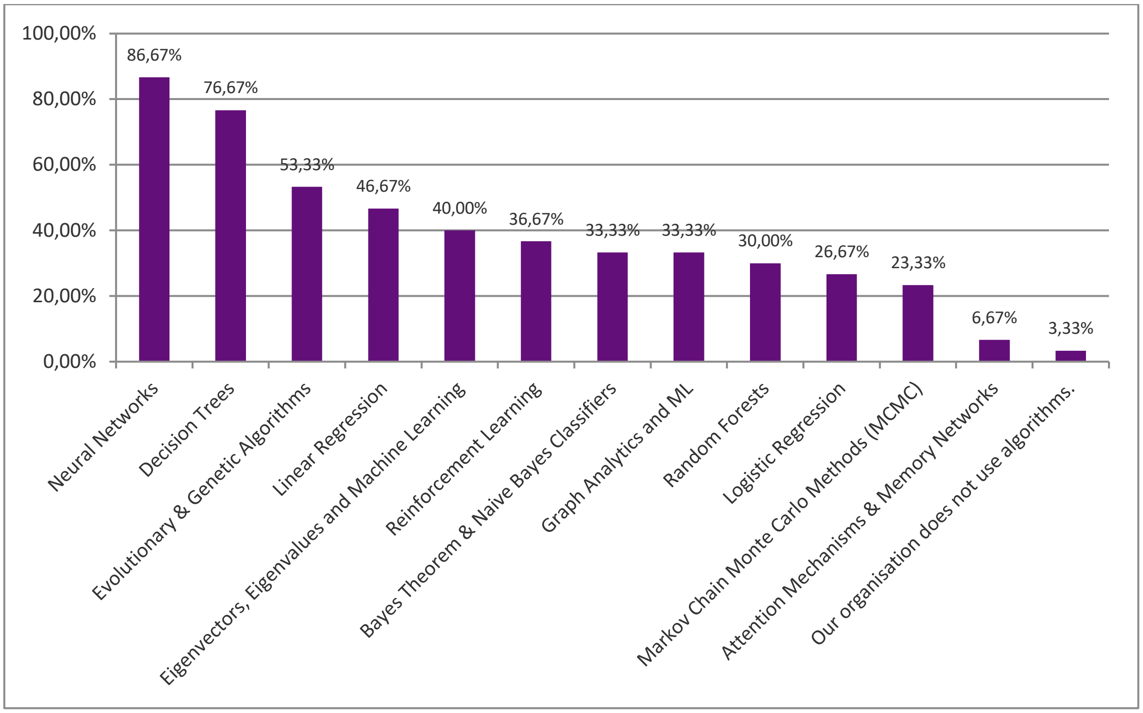


Figure 7: Distribution of algorithms used by AI technology solution providers

Algorithms are of course complex functions and represent an important intellectual property of technology solutions providers. Therefore, development in the form of algorithm training is an important task for technology providers to ensure their continuous competitiveness Depending on the size of providers' organisations, these tasks cannot be fulfilled completely in-house. However, asked about their approach towards training of algorithms, two-thirds of responding solution provider state to rely on internal resources to perform algorithm training. Many solution providers do not work completely on their own but rely co-developing their algorithms with industrial partners (46 %). A comparable share uses publicly available data to train their algorithms (40 %). Other, less commonly applied strategies include using customer data to train AI algorithms from the ground up or relying on network partners for support.

An important aspect when developing machine learning algorithms is to ensure that their decisions do not accidentally discriminate against certain groups of people. A known example is an algorithm that was used in human resources departments that discriminated against women when hiring, because of the input data being historic hiring statistics featuring predominantly male workers. To prevent this from happening, developers are asked to build their algorithms in a way that features protection mechanisms. However, when asked whether they are making sure that their algorithms are protected from acting in a discriminating way, responses from AI solution providers showed strongly mixed results. While almost 47 % stated to make sure this does not happen, an almost identically large share of 40 % said they would not take action to prevent discrimination from occurring, while little over 13 % of solution providers stated to not have known algorithm-based discrimination was an issue after all (see Figure 8).

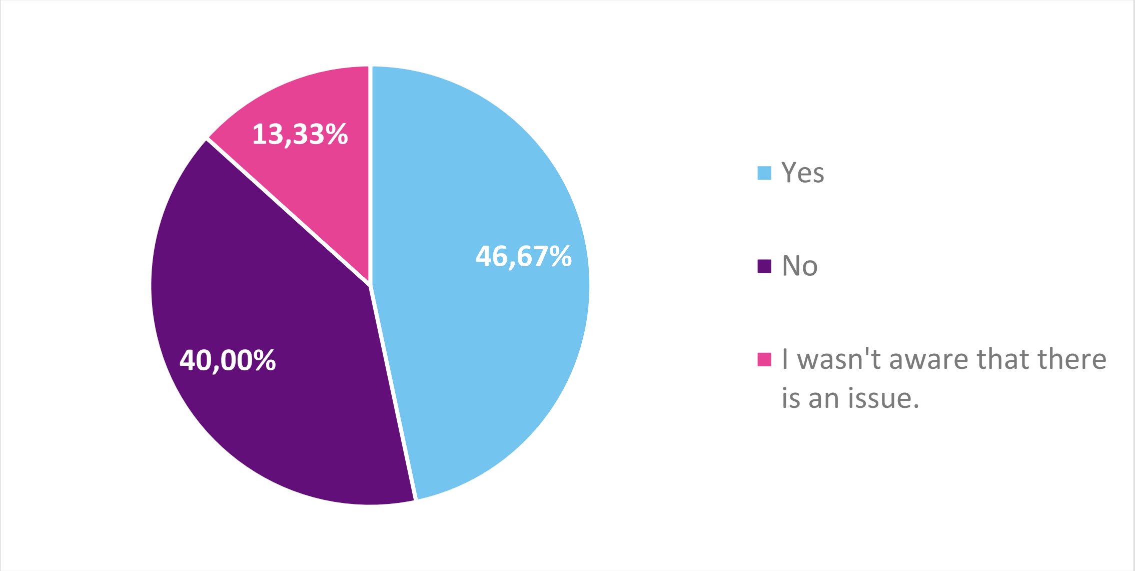


Figure 8: Handling of algorithm-based discrimination by technology solution providers

4. Outlook

The results gained by the survey show that the issues of data and algorithms feature plenty misunderstandings in the communication between organisations looking to adopt AI solutions and organisations offering AI technology solutions. In addition to general lack of AI-related expertise among organisation's workforces, limited capacities and funding, these troubles contribute to existing struggles and decelerate the rate at which AI technology is finding its way into daily business operation, at least regarding non-digital-first businesses.

4.1 Contribution to Knowledge Innovation Triangle

The survey results show that educational activities should take differences in technical background into account to make sure communication between adopters and providers can occur eye-to-eye. Especially educational content to help parties evaluate data quality, availability, and security in a common way would be very helpful. Furthermore, since the responses about algorithm-based discrimination showed mixed results, educational programs directed towards developers of algorithms need highlight the social responsibility attached to their creation.

The survey results also shows that start-ups and scale-ups looking offering AI technology solutions need to make sure their communication towards (potential) clients is adapted accordingly, to avoid misunderstandings or even mistrust in the long-run.

Due to these findings, as well as insights from other cross-KIC activities, the 2021 proposal to establish a platform to support communication and exchange between organisations looking to adopt AI technology solutions and those offering those services emerged. The platform is intended to be approachable for organisations interested in adopting AI solutions. By offering a self-assessment tool, which features an easy-to-understand guide to evaluating the data availability, accessibility, integrity, and security situation, a unified common ground for discussion is created. Furthermore, by asking users to provide information on their processes, the potential of implementing AI solutions will be evaluated. Based on this and other information, the organisation will be directed to a person of a suitable KIC, who will be able to help this organisation fulfil their respective needs by connecting them with players from the three KIC pillars:

- Education
  Highlight educational courses for organisation employees or find suitable graduates from EIT educational programs
- Innovation
  Connect organisations to innovation activities suitable to their cause
- Business Creation
  Connect organisations to start-ups working on customised solutions

4.2 Dissemination / communication activities

The survey and its results were communicated in online workshop as well as within the project's major dissemination event. Furthermore, the survey results will be featured in an explanatory video that's currently being produced and will be published online early next year.