Methodology for quantitative data analysis

In order to analyze the allocation of funding within the Horizon program, we used the *project, organization* and *euroSciVoc* datasets provided on the open-source European Data portal[1]. These datasets were downloaded on 2 September 2024, and which point not all the funding for the Horizon had been allocated yet. For each project in the *project* dataset, its keywords as given on the project's CORDIS webpage[2] were scraped on 4 November 2024. Furthermore, the fields of science according to the European Science Vocabulary for each project were retrieved from the *euroSciVoc* dataset. The union of the CORDIS keywords and the EuroSciVoc fields of science are here referred to as a project's keywords.

These keywords were used to group projects into categories. First, we ranked the keywords according to the number of projects to which they had been assigned and then manually labeled often occurring keywords into "technology" (e.g., *AI*, *Sensors*) and "strategy" (e.g., *Trust*, *Security*). It is worth noting that labeling was done by people primarily with a background in digital technologies, meaning that keywords related to e.g. biology might not be as thoroughly labeled as, for instance, those related to AI. Based on these manual labels, some very simple phrase-based classifications were carried out. Through analyzing the results of this first classification, we found that no clear separation could be made between technologies and strategies. *Health* could be both a strategy as a technology, for example. Because of this, we decided not to group the keyword categories into technologies or strategies. Based on the analysis, we also heuristically improved the keyword labels. The keyword "ecosystem" was removed, for example, because it is used in different contexts with very differing meanings: in computing contexts, it refers to *software ecosystems* while in biology contexts it refers to ecosystems in the ecological sense.

The classification method that we finally used is a combination of phrase-based matching and word-sense disambiguation. Given the manually compiled labels, a keyword is assigned categories as follows. First, it is checked whether a completely stripped (such that it is reduced to only its lowercase alphanumeric characters) version of the keyword has a direct match in the stripped versions of the labelled keywords. If this is the case, the keyword is assigned that labelled keywords' category. If no such match is found, the tokens (e.g. words) in the keyword are matched with the labelled keywords. If a labelled keyword matches with a token, its category is added to the keyword's categories. The same is done for any subset of the keyword, meaning labelled keywords are also matched if they only form part of a token of the keyword,

[1] https://data.europa.eu/data/datasets/cordis-eu-research-projects-under-horizon-europe-2021-2027?locale=en
[2] https://cordis.europa.eu/

but for this, labelled keywords containing less than 4 characters are skipped as to prevent meaningless matches (e.g. *AI* in *container*). If token matching did not lead to any categories being assigned to the keyword, word sense disambiguation is carried out using the *Natural Language Toolkit*[3]. The synsets of the keyword are compared with the synsets of the labelled keywords and if there is a WuPalmer[4] similarity greater or equal to 0.95 between any of the synsets of the keyword and any of the synsets of the labelled keywords, the keyword is assumed to match with the labelled keyword and its category is added to the keyword's categories. Word sense disambiguation is only applied if synsets could be assigned to at least ¾ of the keyword's tokens.

After having applied this classification method once, we again heuristically adjusted the keyword labelling. We then found, however, that that keyword labelling contained so many categories that it became very hard to oversee the data. Therefore, we decided to group the categories into categories, subcategories and subsubscategories. The classification was carried out again, and the results were analyzed. The results of this analysis can be found in the *Results* section below.

The code for the steps described above along with the code used to obtain the results in the section *Results* can be found on GitHub[5].

Limitations

The CORDIS datasets were downloaded on 2 September 2024. The sum of ecMaxContribution across projects in that project dataset is €36.8 billion, which is substantially lower than the €95.5 billion allocated for the Horizon program[6]. This indicates that at that time, not all the Horizon money had been allocated to a project yet. While this means that the data analysis here is limited, the data analysis does give a representative indication of the money allocated from 2021 - September 2024.

As mentioned above, the manual classification for this project was done by people with a digital technology background. This leads to keywords related to digital technologies being classified in a much more fine-grained manner than keywords in, e.g., the health field. This might skew the results of the data analysis, but this has a limited impact on the overall analysis as the analysis is focused on digital technologies.

Furthermore, not all categories have the same number of labelled keywords

[3] https://www.nltk.org/

[4] Zhibiao Wu and Martha Palmer. 1994. "Verb Semantics and Lexical Selection," *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (June 27, 1994): 133-138, https://doi.org/10.3115/981732.981751.

[5] https://github.com/Meret6832/horizon-funding-analysis/

[6] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/programmes/horizon

associated with them. This may skew the data somewhat, as a category with a larger number of keywords has a higher probability of matching a given keyword.

For the word sense disambiguation that is part of the classification method, synsets are assigned to the keyword that is to be classified and the labelled keywords. A given word can have many synsets, the word “cloud” has 14, for example. Which synset may be relevant in the case of that keyword can be determined based on the context, but this context is very limited in the case of keywords. Furthermore, the keywords for any given project may not be related, so other keywords cannot be used as context either. This leads to synsets being included that may not correspond to the intended meaning of a keyword, which lowers the accuracy of the classification method.

Results

Organizations

AcitvityType

Organizations that receive funding under the Horizon program are assigned an activityType, referring to the type of organization. In Table 1 and Figure 1, the distribution of the total netEcContribution of the Horizon program over these activityTypes is shown. The netEcContribution is the total money an organization keeps for a project. It does not include the money distributed to other parties, but does include such money that has been received from other organisations. Here, we can see that Higher or Secondary Education Education Estabilishments receive the most money, followed by Private For-Profit Entities and Research Organisations.

ActivityType	**Meaning**	**netEcContribution (€)**	**% of total netEcContribution**
HES	Higher or Secondary Education Establishments	12.9 billion	35.03%
PRC	Private For-Profit Entities	10.3 billion	27.98%
PUB	Public Bodies	1.30 billion	3.52%
REC	Research Organisations	10.1 billion	27.45%
OTH	Other (including associations, membership organisations and NGOs)	2.2 billion	5.95%

Table 1: Distribution of netEcContribution across activityTypes within the Horizon program.

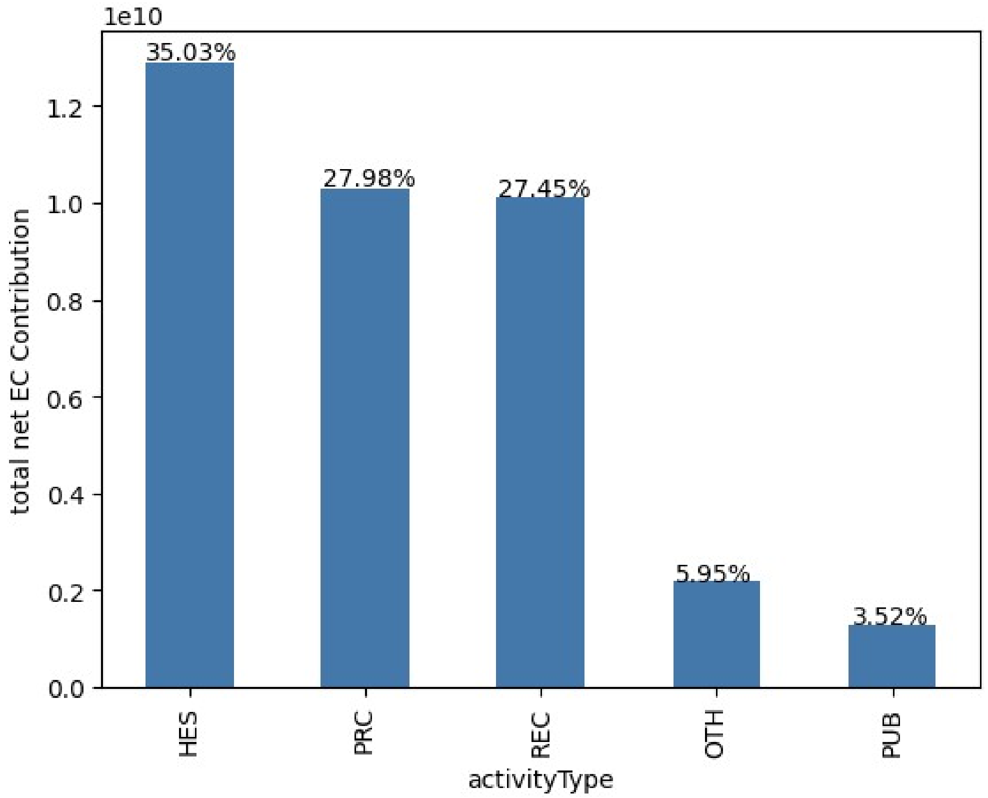

Figure 1: Total netEcContribution and the percentage this forms of the total Horizon program funding per activityType.

In Figure 2, the distribution of netEcContribution over activityTypes per Horizon cluster is shown. Here, we can see that this distribution varies greatly across clusters. More than half of the netEcContribution goes to Private For-Profit Entities in clusters ERC-SJI (56.3%, Science Journalism Initiative[7]), HORIZON-EIC (62.4%, EIC Accelerator[8]), HORIZON-EIT (61.5%, European Institute of innovation and Technology[9]), HORIZON-ER (75.8%, EU Rail[10]), HORIZON-EUSPA (50.4%, EU Space[11]), HORIZON-JTI (67.6%, sustainability in the hydrogen value chain[12]), HORIZON JU (53.8%, Joint Undertaking[13]), HORIZON_KDT (63.4%, Key Digital Technologies[14]), and HORIZON-

[7] https://erc.europa.eu/apply-grant/science-journalism-initiative
[8] https://eic.ec.europa.eu/eic-funding-opportunities/eic-accelerator_en
[9] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/european-institute-innovation-and-technology-eit_en
[10] https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/europes-rail-joint-undertaking_en
[11] https://www.euspa.europa.eu/opportunities/horizon-europe
[12] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-jti-cleanh2-2024-05-01
[13] https://www.welcomeurope.com/en/the-list-of-our-calls-projects/heu-ju-research-and-innovation-actions-supporting-the-global-health-edctp3-joint-undertaking-2024/
[14] https://www.era-learn.eu/network-information/networks/key-digital-technologies

SESAR (68.9%, air traffic management[15]). Most of these clusters are related to digital technologies and are focused on public-private partnerships.

Most-funded organizations

The ten organisations that receive the most funding across all clusters in the Horizon program are shown in Figure 3. Most of these organizations are Research Organisations or Education Establishments. EIT Raw Materials GMBH[16] and EIT KIC URBAN MOBILITY SL[17] are the only Private For-Profit organizations. Both are in initiatives by the European Institute of Innovation and Technology (EIT) and partner with private and public organizations. This shows the limit of these datasets: the organization given in a dataset may not be the ultimate organization to which money has been given. EIT FOOD[18], which is also part of EIT and COST ASSOCIATION, a "funding organisation for research and innovation networks"[19] are classified as Other.

[15] https://www.sesarju.eu/discover-sesar
[16] https://www.eitdeeptechtalent.eu/the-pledge/meet-the-pledgers/eit-raw-materials/
[17] https://www.eiturbanmobility.eu/who-we-are/about-us/
[18] https://www.eitfood.eu/about-us
[19] https://www.cost.eu/

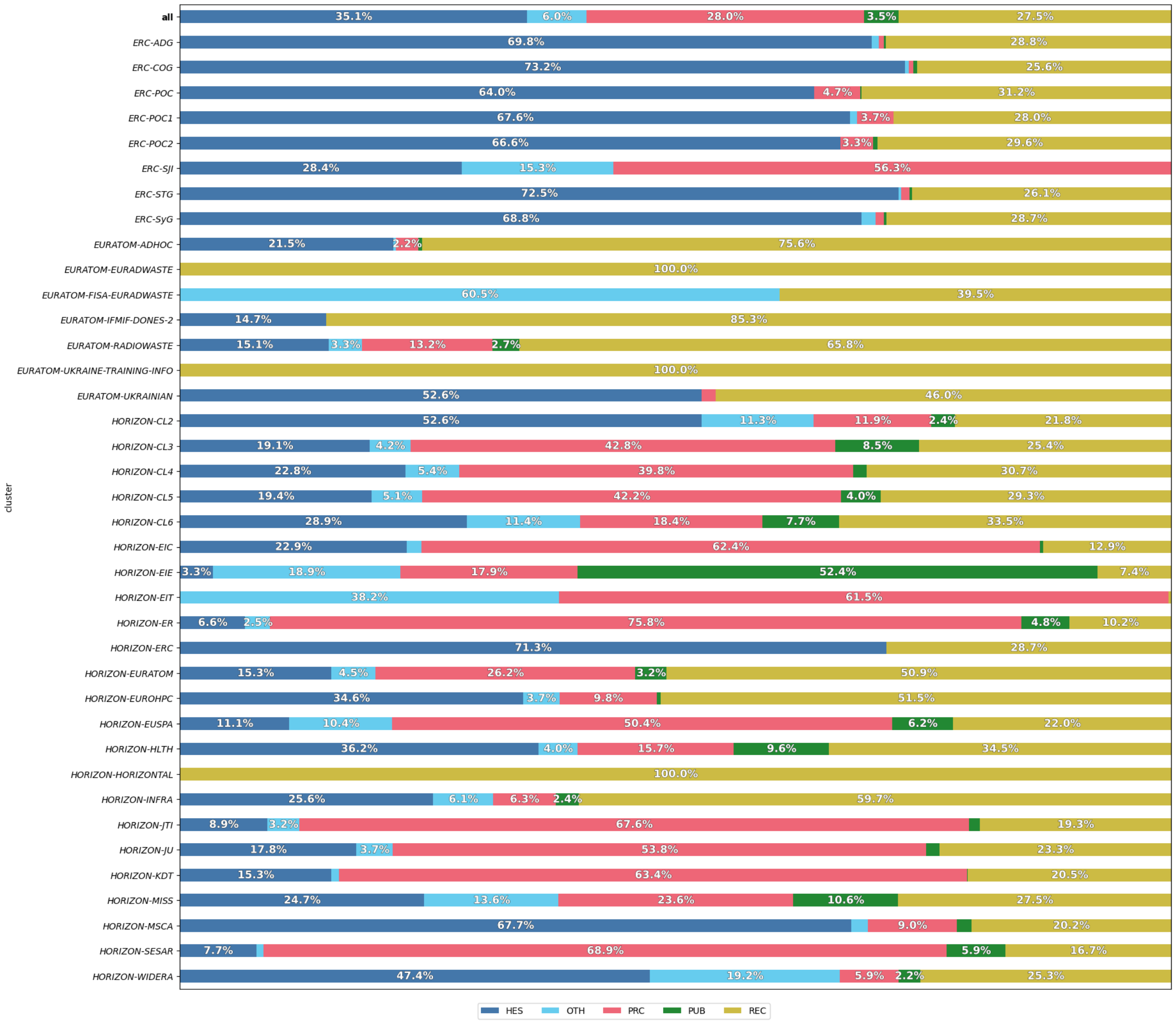

Figure 2: Distribution of netEcContribution over acitivityTypes per Horizon Cluster.

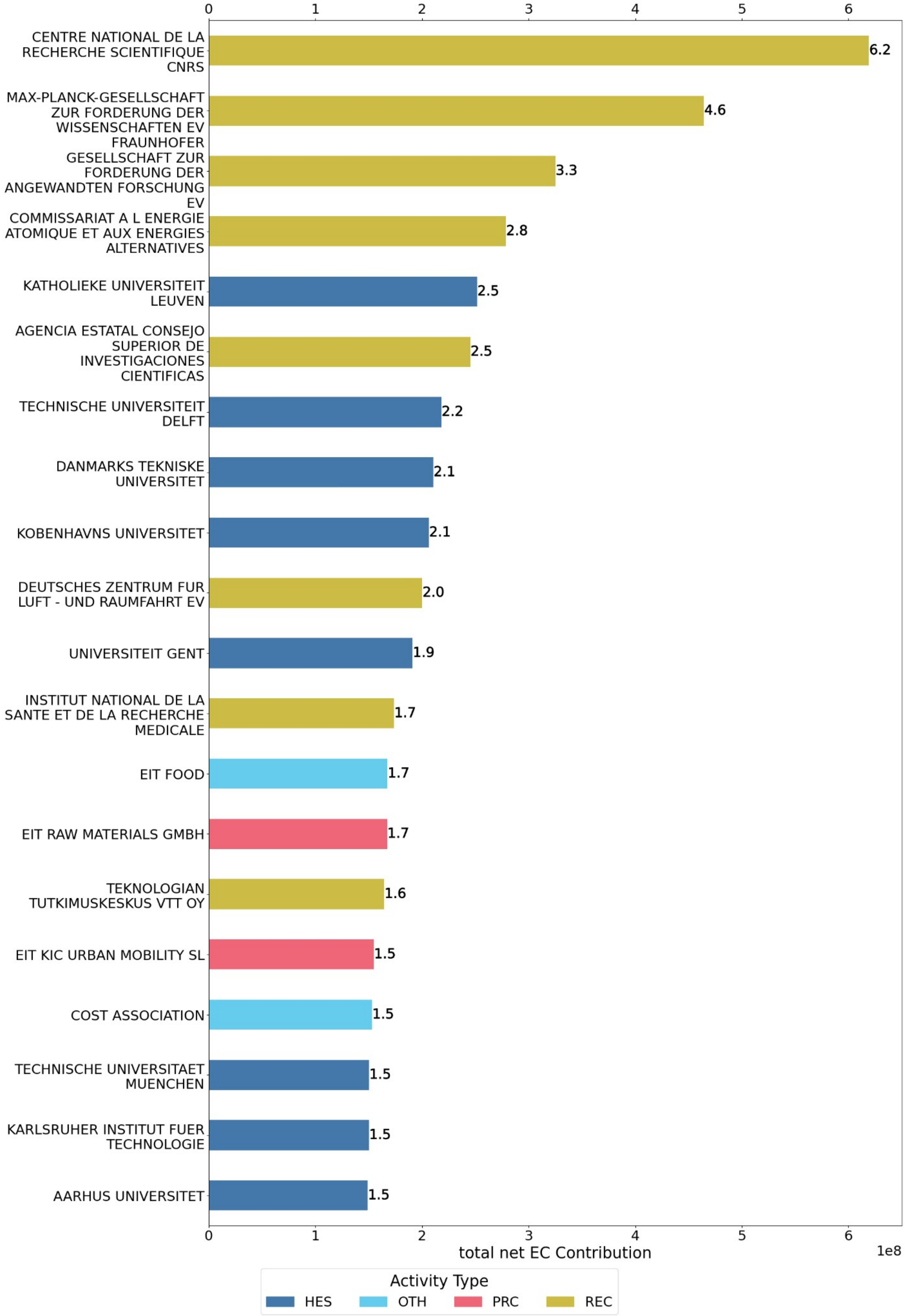

Figure 3: NetEcContribution of the ten organizations receiving the most funding in the Horizon program, amounts shown next to the bars are in € 10^8 (hundred million).

Countries

In Figure 4, the ten countries that receive the most funding in the Horizon program are shown. It is not the case that the governments of these countries receive funding, but organizations that are based in these countries do

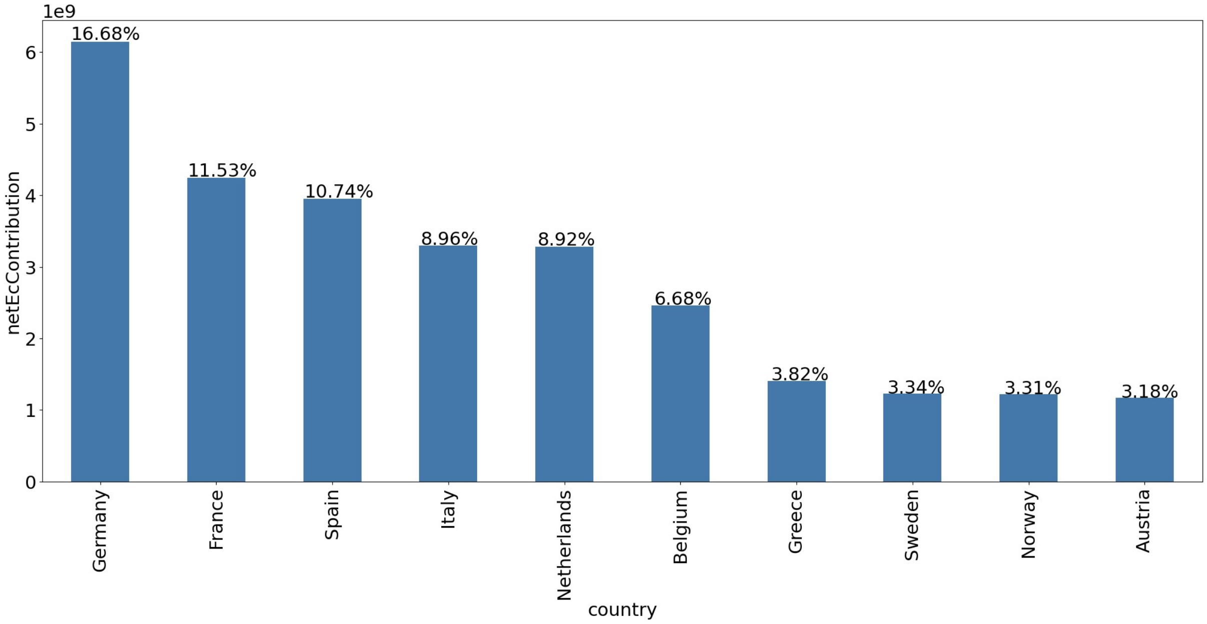

Figure 4: NetEcContribution for the ten countries receiving the most funding in the Horizon program.

All of the ten countries that receive the most funding are part of the European Union (and located in Europe). Indeed, 91.87% of Horizon funding goes to organizations in the EU. In Figure 5, we can see that the vast majority (95.67%) of Horizon funding goes to organizations located in Europe. The ten non-EU countries that receive the most funding are displayed in Figure 6. Only three of these ten countries (Israel, South Africa and the United States) are not (partially) located in Europe.

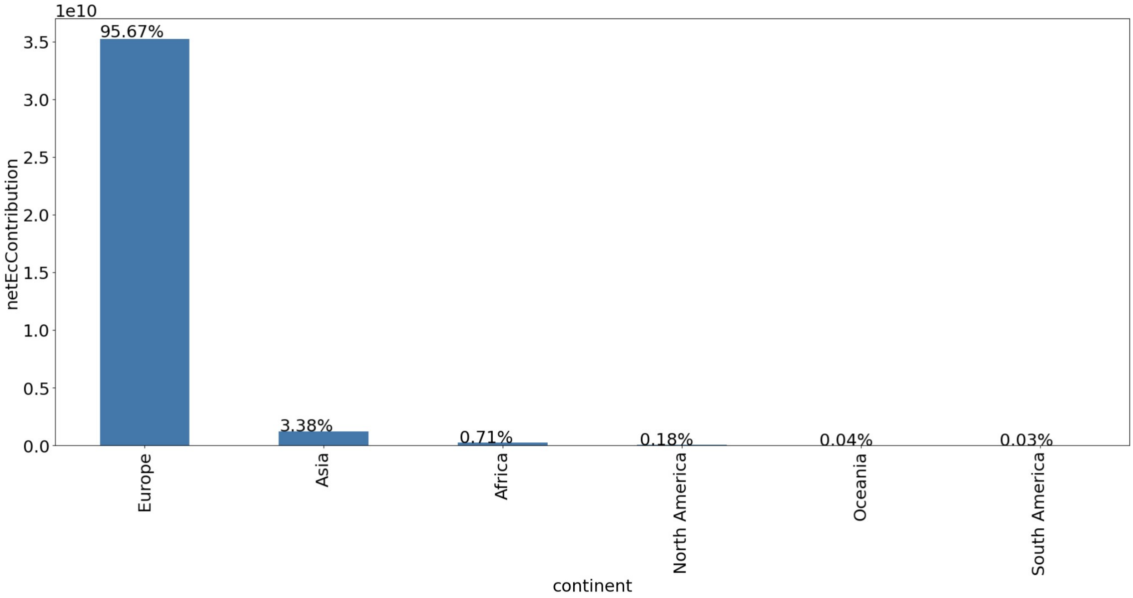

Figure 5: Distribution of netEcContribution of the Horizon program across continents.

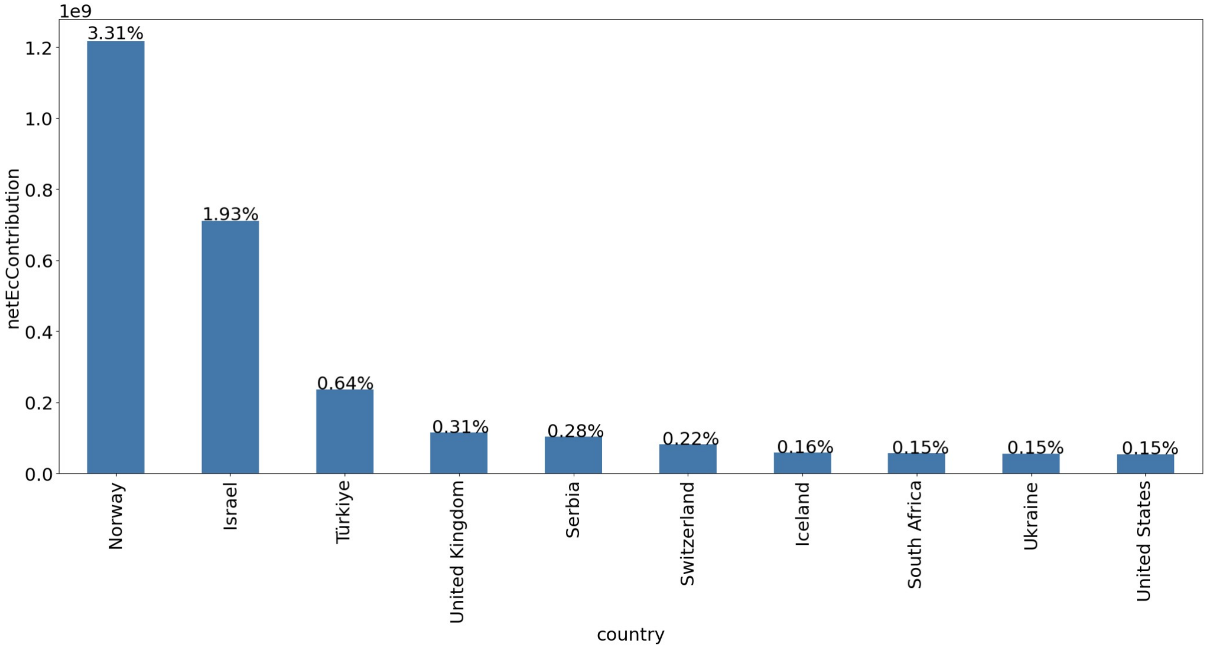

Figure 6: NetEcContribution of the ten non-EU countries that receive the most funding in the Horizon program.

Categories

In this section, we analyze what topics are funded under the Horizon program through the results of the classification discussed in the *Methodology* section. An overview of some general statistics for these categories can be found in the *overviewCategories* table on GitHub[20].

In Figure 7, the 25 categories that receive the most funding in the Horizon program are shown. It is important to note that one project can be assigned to multiple categories. For each of these categories, the project's ecMaxContribution is then counted for that category. The sum of the ecMaxContribution across all categories here is therefore greater than the total ecMaxContribution. Digital Technologies, Sustainability and Health are the categories that are funded most overall. Raw Materials, Aerospace and Explainability obtain the most funding per project, on the other hand (see Figure 8).

[20] https://github.com/Meret6832/horizon-funding-analysis/blob/main/overviewCategories.csv

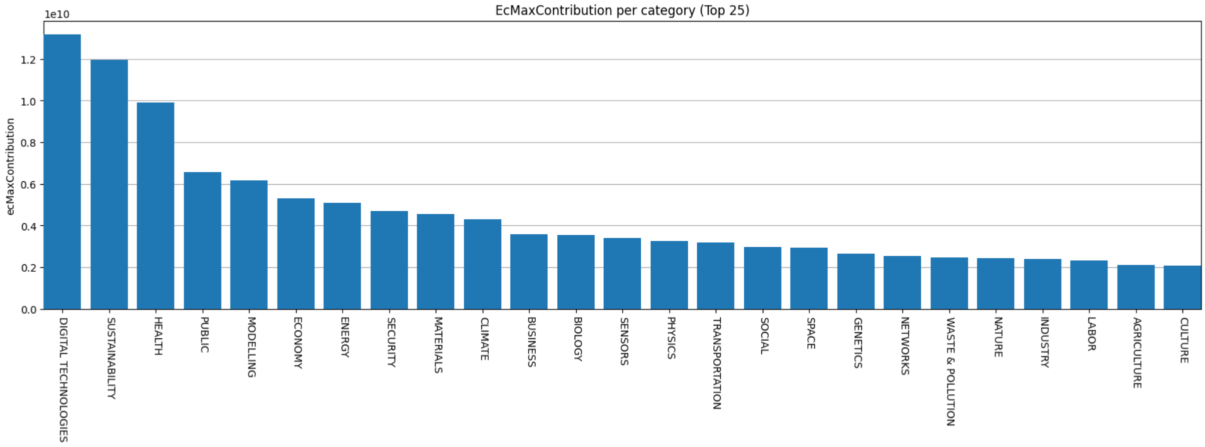

Figure 7: EcMaxContribution of the 25 most funded categories.

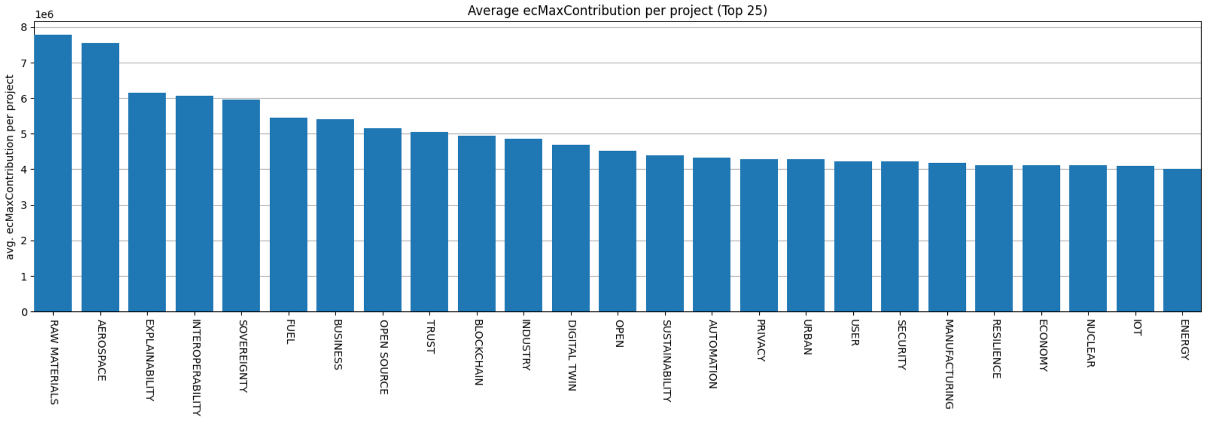

Figure 8: Average ecMaxContribution per project for the 25 categories with the highest average ecMaxContribution per project.

Digital Technologies

In this section, we zoom in on the Digital Technologies category. Within this category, projects related to AI receive the most funding, followed by projects related to Data and Computing. Most projects fall in at least one of the categories, as the ecMaxContribution of the "other" grouping is quite small.

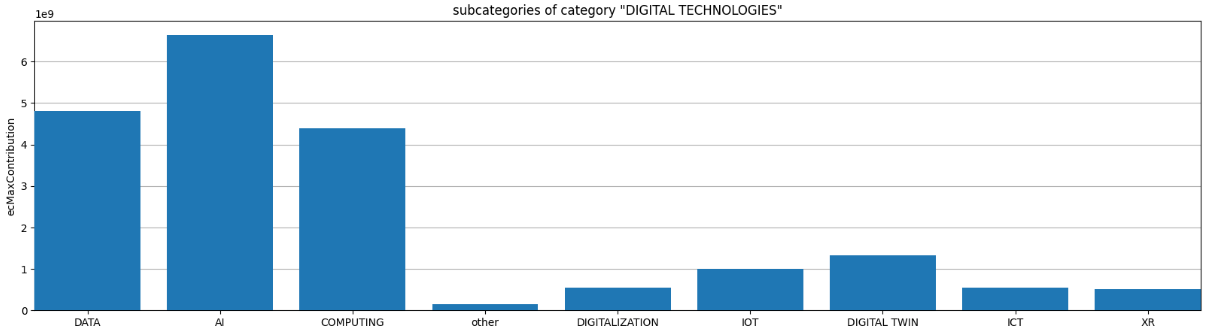

Figure 9: EcMaxContribution of the Digital Technologies subcategories. Projects that fall under Digital Technologies but not any of the subcategories are grouped under "other".

When we zoom in further on AI (see Figure 10), we can see that, of AI’s subsubcategories, Machine Learning receives the most funding. Generative AI and LLM receive the least of these subsubcategories.

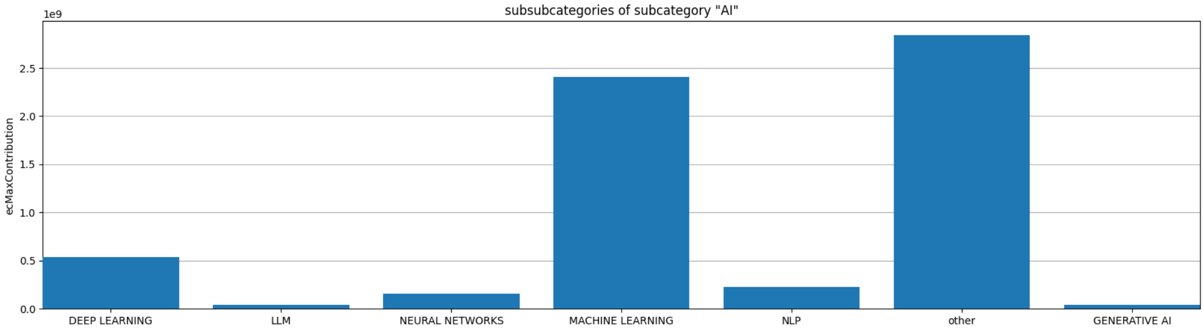

Figure 10: EcMaxContribution of the AI subsubcategories. Projects that fall under Digital Technologies but not any of the subcategories are grouped under “other”.

For the subcategory Computing, the most funding goes to the subsubcategory Software. Supercomputing is receives the least ecMaxContribution (see Figure 11).

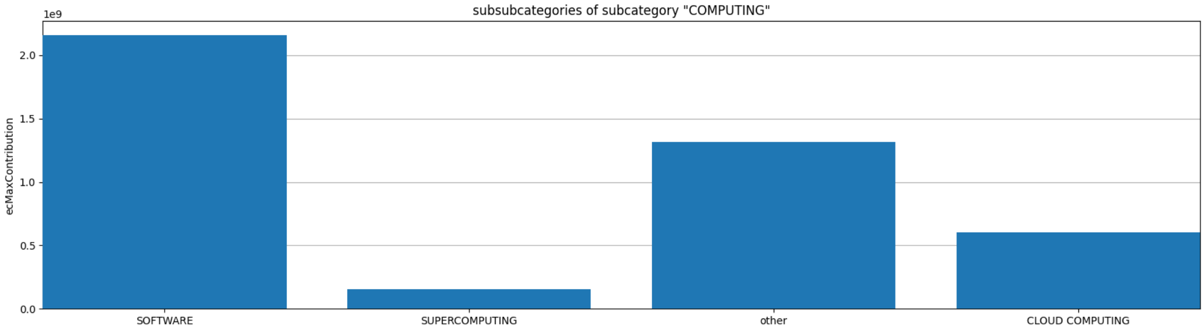

Figure 11: EcMaxContribution of the Computing subsubcategories. Projects that fall under Digital Technologies but not any of the subcategories are grouped under “other”.

We can also look at the co-occurrences of categories, i.e. the categories that occur most together with another category within one project. In Figure 12, the 15 categories that occur with the most funding relative to the other category’s funding are displayed. Here, we can see that approximately 70% of funding in the Robotics and Privacy categories goes to projects that are also related to AI. This is also the case for more than 50% of categories in the Trust and Law Enforcement categories.

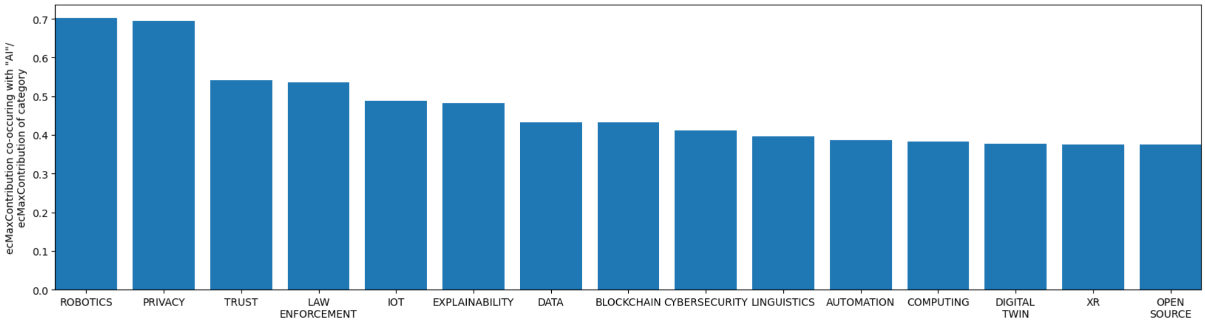

Figure 12: Fraction of ecMaxContribution that goes to projects that are also related to AI for the 15 categories, subcategories and subsubcategories for which this fraction is the highest across the Horizon program. Super- and subcategories of AI have been excluded.

For Sovereignty, the most co-occurring categories are shown in Figure 13. Here, we can see that Sovereignty co-occurs most with projects related to User, Interoperability and Privacy, but that these co-occurrences less than 10% of those categories' funding.

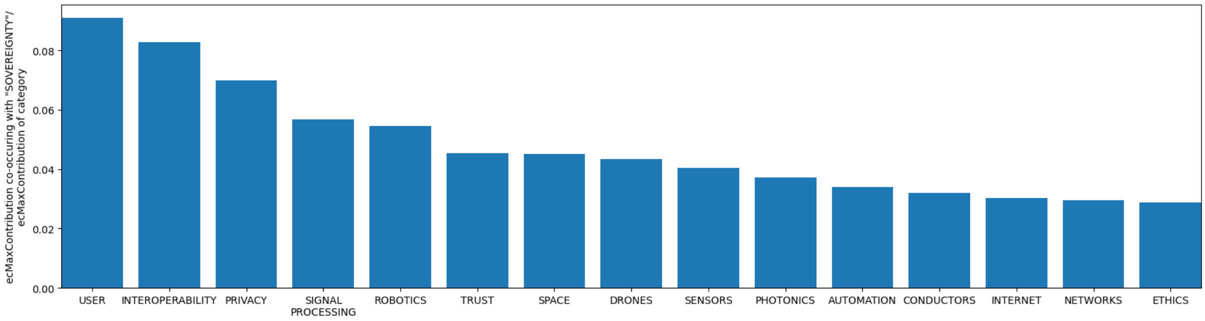

Figure 13: Fraction of ecMaxContribution that goes to projects that are also related to Sovereignty for the 15 categories, subcategories and subsubcategories for which this fraction is the highest across the Horizon program.

Finally, the co-occurrences of the Military category are shown in Figure 14. Maritime, Bioinformatics and are the most co-occurring categories and more than 10% of their funding goes to projects that are also related to Military. After these three, Surveillance is the most co-occurring category.

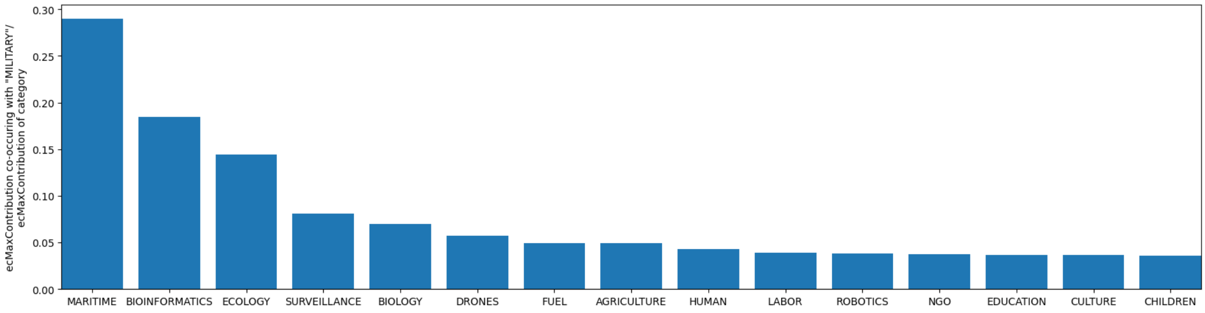

Figure 14: Fraction of ecMaxContribution that goes to projects that are also related to

Military for the 15 categories, subcategories and subsubcategories for which this fraction is the highest across the Horizon program.

AI across clusters

In Figure 15, the percentage of the total ecMaxContribution of a cluster that goes to projects related to AI is shown. Although HORIZON-CL4 (*Digital, Industry and Space*) is the cluster dedicated to digital technologies, HORIZON-CL3 (*Civil Security for Society*) and HORIZON-EUSPA (*Space*) are the clusters with the greatest percentage of ecMaxContribution going to projects related to AI.

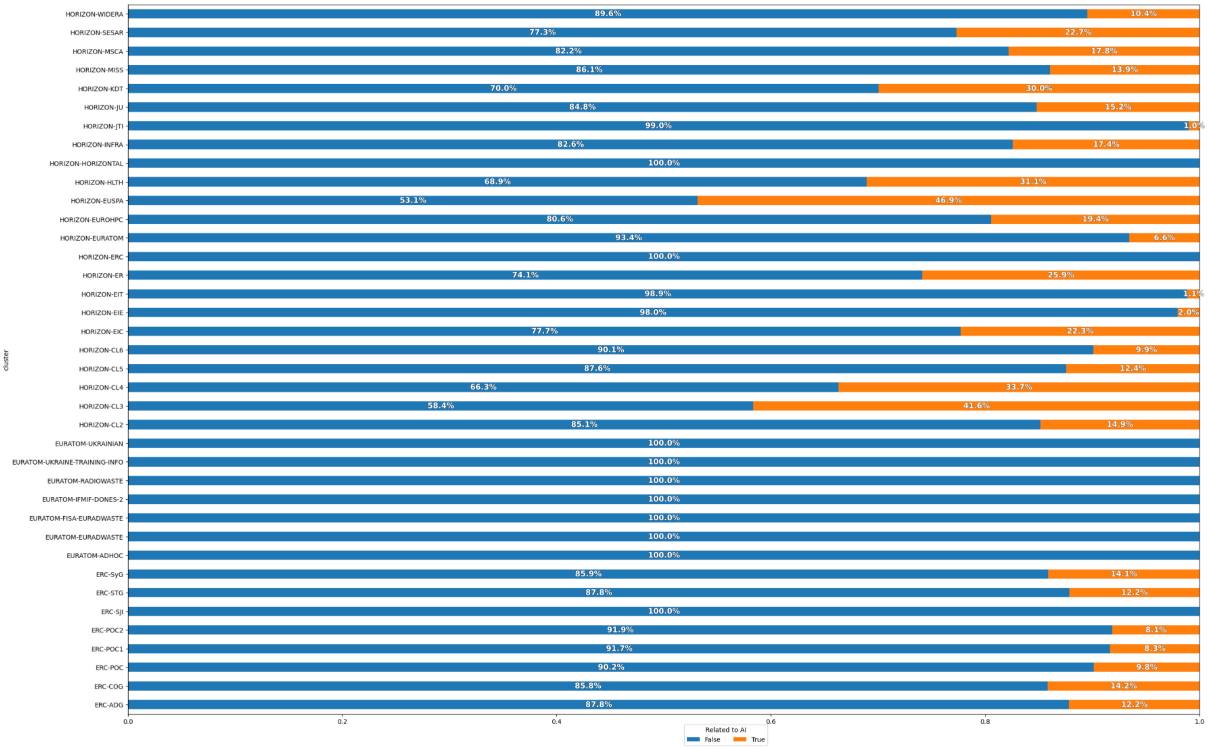

Figure 15: Percentage of ecMaxContribution that goes to projects related to AI (orange) for each cluster in the Horizon program.

ActivityTypes per Category

Finally, we can also look at the distribution of activityTypes within certain categories.

Within the Military category, on the other hand, most funding goes to Research Organisations (see Figure 16). On the other hand, within Digital Technologies, the most funded category, most funding goes to Private For-Profit Entities (see Figure 17). This is also the case for the Digital Technologies subcategory AI (see Figure 18)

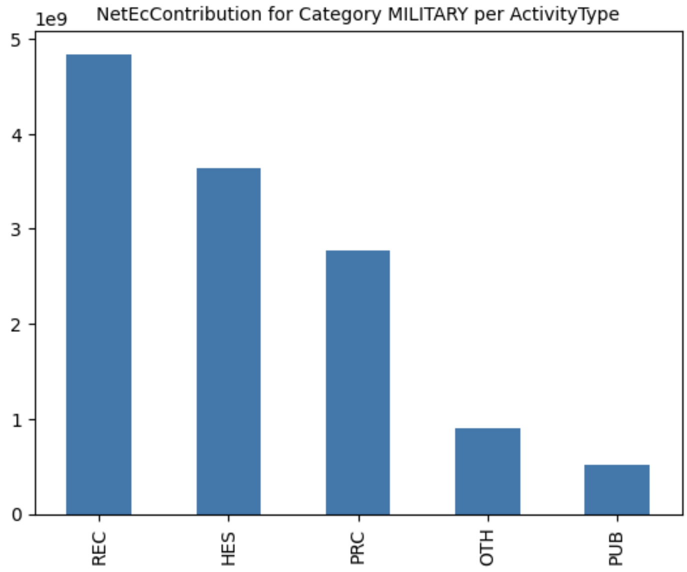

Figure 16: NetEcContribution distribution across activityTypes within the Military category.

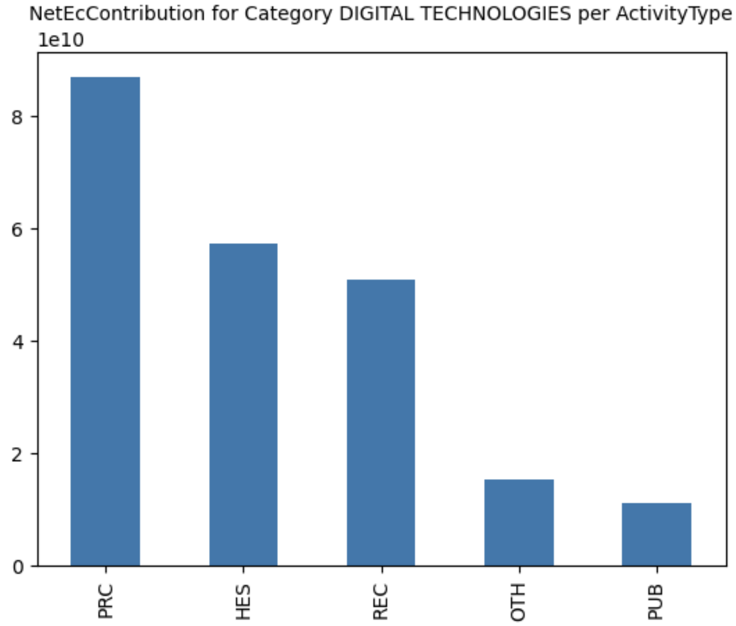

Figure 17: NetEcContribution distribution across activityTypes within the Digital Technologies category.

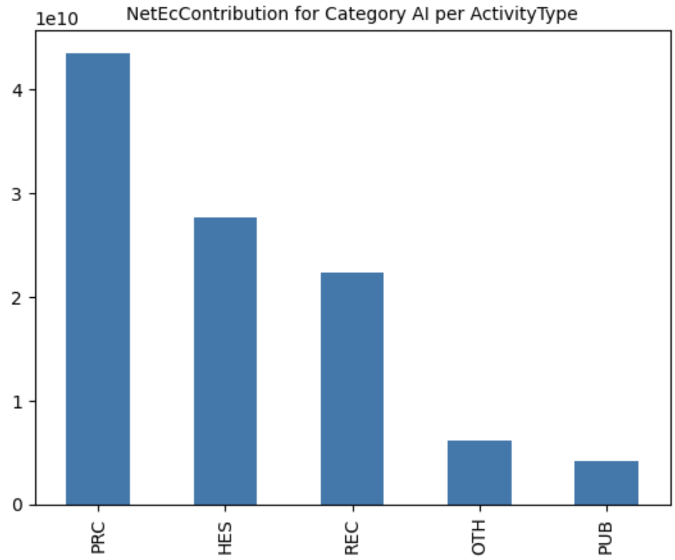

Figure 18: NetEcContribution distribution across activityTypes within the AI subcategory.

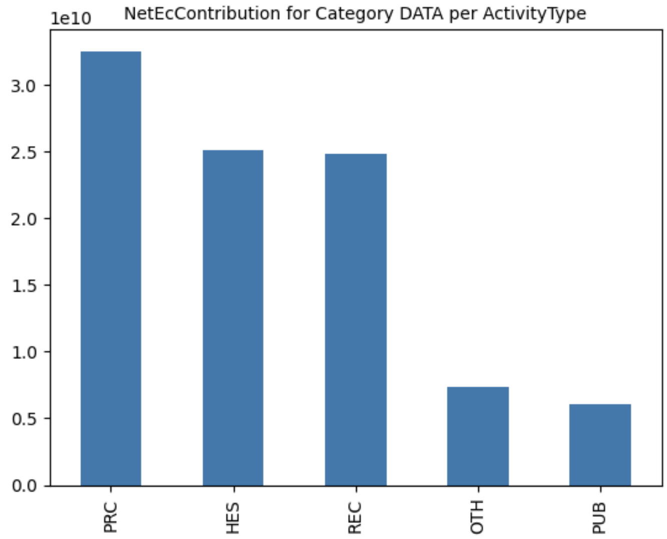

Figure 19: NetEcContribution distribution across activityTypes within the Data subcategory.