EXTRACTING THE SIGNAL FROM THE NOISE - THE ROLE OF ARTIFICIAL INTELLIGENCE IN THE ANALYSIS OF SAFEGUARDS RELEVANT INFORMATION

P. SCHNEEWEISS
IAEA
Vienna, Austria
Email: p.schneeweiss@iaea.org

T. STOJADINOVIC
IAEA
Vienna, Austria
Email: t.stojadinovic@iaea.org

Z. ABBALI
IAEA
Vienna, Austria
Email: z.abbali@iaea.org

N. GILLARD
IAEA
Vienna, Austria
Email: n.gillard@iaea.org

M. NITECKI
IAEA
Vienna, Austria
Email: m.nitecki@iaea.org

Abstract

Through safeguards, the IAEA is able to provide credible assurances that States are honouring their international obligations to use nuclear material only for peaceful purposes. Therefore, the IAEA collects, processes and evaluates safeguards relevant information about a State from three sources: State-provided information, information from IAEA safeguards activities and other relevant information (for example, from open sources, satellite imagery or information provided by third parties). The large and growing volume and variety of available information underlines the need for new approaches to support IAEA analysts and IAEA subject matter experts in their assessment of safeguards relevant information. By leveraging recent breakthroughs as well as long-established methods in the field of natural language processing, our work shows how data-science-based approaches are being applied to the nuclear domain to extract the "signal" from the "noise", thus strengthening the effectiveness and improving the efficiency of IAEA safeguards.

1. INTRODUCTION

The collection and evaluation of all safeguards relevant information is one of the fundamental processes of IAEA safeguards implementation. The rapid increase in availability of large amounts of data across media types calls for new approaches to assist analysts in spotting "the signal in the noise". Artificial Intelligence (AI) and Machine Learning (ML) offer substantial improvements to effectiveness and efficiency by supporting safeguards analysts in their assessments, and allowing them to focus on tasks which require more complex analysis.

There is a broad consensus that digital data increases each year in an exponential manner [1], [2]. Even though this exponential trend has flattened off in some safeguards relevant sources [3], the number and variety of sources, websites and media types are increasing rapidly and present new challenges for the IAEA. Against a backdrop of resource constraints and a vast increase in potentially relevant information, the IAEA must balance the breadth of its dataset against the resources required for collection and analysis. To overcome this challenge, new technologies are being explored that might assist analysts in their daily work and ensure that relevant information is spotted.

The high maturity level of modern AI algorithms has brought great advances in various domains, in particular in applications that include unstructured data. In comparison to traditional machine learning algorithms, deep learning models are able to better understand semantic and contextual information in images and text. The success of those algorithms comes primarily from the availability of large data and computing power, as well as from some algorithmic innovations. [4] It is noticeable that the availability of large amounts of data, while posing a challenge to traditional analysis tasks, was the origin of a remarkable evolution in machine learning models. This is especially true in the field of natural language processing with transfer learning and domain adaptation, where what has been learned in one setting is exploited to improve generalization in another setting.

The guiding research question of the paper is: what role do modern algorithms play in supporting the analysis of safeguards relevant information? To address this question, a selection of current use-cases is depicted and first results are discussed to evaluate the potential of AI and data science for IAEA Department of Safeguards.

2. CHALLENGES OF ANALYSING SAFEGUARDS RELEVANT INFORMATION AND CONSISTENCY ANALYSIS

Open source (OS) information analysis is an essential element in the Department of Safeguards' effort to detect possible undeclared nuclear activities and the misuse of declared facilities and material. The Department of Safeguards must identity and collect all safeguards relevant information from open sources, including but not limited to scientific and technical (S&T) publications, news, government records, trade information and social media. Even though the exponential trend in the digital data increase has flattened off in some safeguards relevant sources [3], the number and variety of sources, websites and media types are increasing rapidly and pose new challenges to the IAEA.

OS analysts routinely employ simple keyword searches or more complex Boolean expressions when querying databases such as Web of Science, Scopus and Science Direct, or search engines like Google and Google Scholar. Simple keyword searches can negatively affect productivity by returning irrelevant results (false positives), whereas high-complex queries require constant refinement and can hinder the discovery of new insights (false negatives).

Those limitations underline the need for semantic search, which takes into consideration the analyst's intent for contextual meaning in search terms. This approach allows as comprehensive and effective collection as possible, avoiding missing important elements available to us while also not drowning the Department with irrelevant information. [5]

No open source information is acted upon in the Department before validation by those with appropriate expertise, and then integration with all safeguards relevant information available to the IAEA. Together with information collected from in-field activities, open source information is used to assess State-provided information; in particular how correct and complete State declarations are.

An example of seeking effectiveness and efficiency improvement through fit-for-purpose AI is related to the fact that States that have signed an additional protocol (AP) declare relevant research activities as part of articles 2.a(i), 2.a(x) and 2.b(i). To date, reviewing those AP declarations involved a high degree of manual work. All States declare their research in different levels of detail, with different approaches in terms of already declared research. As such, identifying research that is new compared to previous years can be a time-consuming task.

Consistency analysis between different sources of information is essential for prioritising the need for additional analysis, and any subsequent engagement with the State or follow up in-field activities. The process of consistency analysis addresses the question of whether the validated open source information is consistent with the State-provided declarations (see Figure 1).

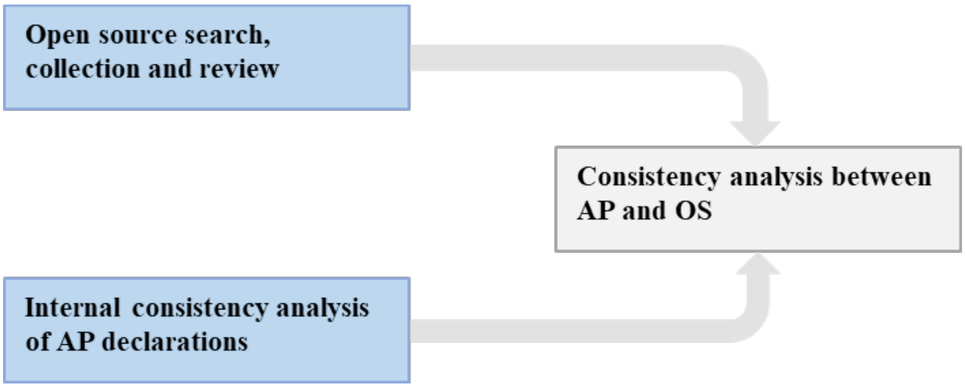

Open source search, collection and review
Consistency analysis between AP and OS
Internal consistency analysis of AP declarations

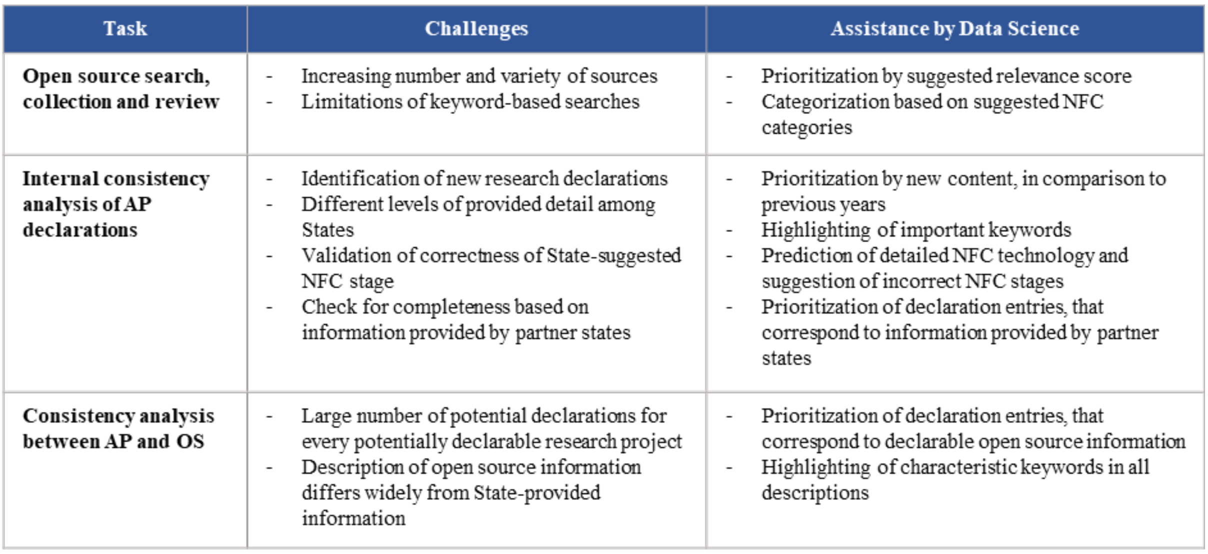
Figure 1: Overview of validation workflow and consistency analysis

There are several challenges in this context which make the consistency analysis of unstructured open source information a highly resource-intensive process. Validated and relevant open source information for a specific State and year can contain several dozens of documents, and declarations made under articles 2.a(i), 2.a(x) and 2.b(i) of the AP can contain up to hundreds of entries of declared research. Declarations from previous years are considered during consistency analysis, and these characteristics can quickly lead to a large number of possible combinations of declarable research and provided declarations. A further challenge is that the description of research projects declared by the State can differ widely from the available open source information. Therefore, common techniques to search inside State declarations for declarable research are full-text searches by characteristic keywords, such as project names, specific nuclear technologies, organizations or locations.

Analysing safeguards relevant information and consistency analysis are time-consuming processes that involve a high degree of manual work, regardless of the tool or tools used to facilitate the process. Section 3 summarizes concrete data science projects and methods which could demonstrate high potential to assist analysts during their manual work.

3. ASSISTING ANALYSTS WITH AI AND ML

A selection of data science projects, which are already in a productive system or in a prototype phase, have had promising initial results in their performance and demonstrate the potential to increase the effectiveness and efficiency of current Safeguards tasks.



Task	Challenges	Assistance by Data Science
Open source search, collection and review	- Increasing number and variety of sources - Limitations of keyword-based searches	- Prioritization by suggested relevance score - Categorization based on suggested NFC categories
Internal consistency analysis of AP declarations	- Identification of new research declarations - Different levels of provided detail among States - Validation of correctness of State-suggested NFC stage - Check for completeness based on information provided by partner states	- Prioritization by new content, in comparison to previous years - Highlighting of important keywords - Prediction of detailed NFC technology and suggestion of incorrect NFC stages - Prioritization of declaration entries, that correspond to information provided by partner states
Consistency analysis between AP and OS	- Large number of potential declarations for every potentially declarable research project - Description of open source information differs widely from State-provided information	- Prioritization of declaration entries, that correspond to declarable open source information - Highlighting of characteristic keywords in all descriptions

Table 1: Overview of AI and ML assistance in analysis tasks

Table 1 summarizes identified challenges (see section 2) and outlines how data science can assist during aspects of open source information search, collection and review; and review of certain Additional Protocol declarations. The sections 3.1, 3.2 and 3.3 provide more details about the data science methods. All projects are aiming to support Safeguards analysts, who are faced with the challenge to find relevant information in large amounts of unstructured data. Therefore, data science is never used to automate tasks or to take decisions, but rather to provide tools for analysts to support their review activities.

3.1. AI assistance for open source search, collection and review

The landscape of potentially safeguards relevant open source information is crowded and constantly evolving. Analysts working with S&T publications face a growing number of records across established and new online databases, such as Web of Science, Scopus, Dimensions and Microsoft Academics. Those databases are routinely queried using keyword searches, which require constant refinement in order to balance the benefits of capturing as wide a dataset as possible and the challenge of the resource-intensive review that wide searching entails.

An AI-powered tool for ranking the safeguards relevance of S&T publications can support analysts in their assessments by focusing their attention on what matters most. For this purpose, a suggested relevance score is assigned to each publication, using a training corpus of validated publications previously collected by IAEA Safeguards analysts. Such prioritization based on safeguards significance allows a timely follow-up in the continuous support of State evaluation, while also ensuring a more comprehensive and effective process.

A variety of models ranging from simple probabilistic to classical machine learning and state of the art neural networks have been implemented. This iterative process required working closely with highly-skilled Safeguards Information Analysts and subject matter experts with relevant competences, in particular in nuclear fuel cycle technologies. One main challenge was collecting and labelling a sufficient number of safeguards relevant publications, which for S&T publications means relating to an existing or developing component of the nuclear fuel cycle (NFC).

A Naïve Bayes classifier, which is known to work well on short documents and small datasets, was implemented as the baseline model. This classifier is also simple, quick to train and performs better than other linear classifiers such as logistic regression and support vector machines. More recently, large language models such as BERT [6] or GPT have shown to work well at solving tasks with limited data by learning useful representation of text that encodes information across many dimensions. The embeddings generated by those models can be used directly to find similar documents or, alternatively, transformer-based models can be further fine-tuned. In particular, we evaluated an in-house domain adapted DistilBERT model as well as multilingual models like XML-RoBERTa.

For relevance ranking of S&T publications, the models were tested on specific nuclear fuel cycle stages. One test set related to centrifuge enrichment is depicted in Figure 2.

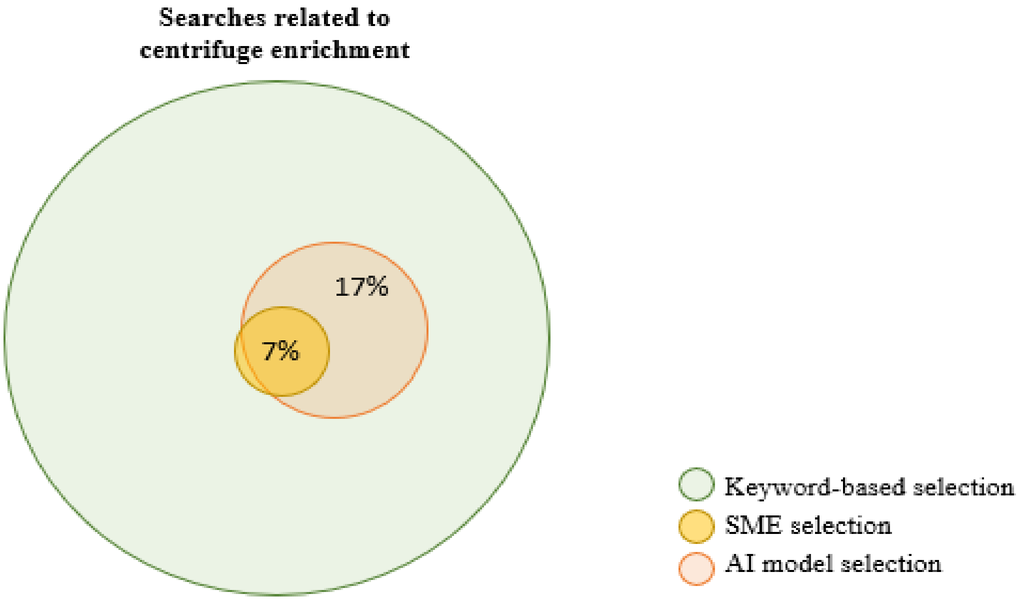


Figure 2: Proof of concept results for computer-assisted ranking of S&T publications

The test set included 148 publications related to centrifuge enrichment of which 11 (7%) were selected by a subject matter expert as being safeguards relevant. Our model selected 26 publications (17%). If the classifier assessment in the case study was selected, reviewers would lose 1 of 11 eligible publications, which was related to laser enrichment rather than centrifuge enrichment, while decreasing the workload by 83%.

The AI model was extended to cover other nuclear fuel cycle categories and a new AI model was trained to cover other types of textual open source information, including websites and news, in order to support ongoing monitoring of all safeguards related activities within a State.

Since we are using a rank-based approach rather than a threshold-based approach, it was decided that an appropriate metric would be the ratio of documents to be retrieved in order to reach a recall level of 0.9. In practice, this means how many documents an analyst has to look at in order to (have assurance that they have seen 90% of the safeguards relevant items. This approach emphasizes highly relevant publications appearing early in the results list, while allowing a margin for documents that are selected but may not be as pertinent.

The AI ranking was applied and showed that it would have been sufficient to look at 20% of the results in order to get 90% of the relevant documents, considerably reducing the workload of the analysts.

This enhanced approach to identify relevant data within the constant flow of information has been integrated into OSIS 2.0 (Open Source Information System 2.0), a system used within the IAEA Department of Safeguards for collecting and processing open source information for safeguards State evaluation. [7]

In addition to the safeguards relevance predictor, a Physical Model classifier was trained, which assigns a suggested corresponding nuclear fuel cycle stage to a document with high relevancy score, as discussed in more detail in Schneeweiss and Stojadinovic (2022). [8] This computer-assisted categorization can be of value in the verification process of additional protocol data, and consistency analysis between AP data and open source information.

3.2. AI assistance for internal consistency analysis of AP declarations

Review of the 2.a(i), 2.a(x) and 2.b(i) AP declarations involves a high degree of manual work. By leveraging methods from natural language processing, those review activities can be supported and made more efficient in various ways.

States' AP declarations relating to the location of NFC related research and development (R&D) are not homogenous in format and substance. To declare ongoing research projects, some States duplicate the declaration entry from the previous year and make minor updates to describe the current status of their research. Other States do not repeatedly declare ongoing research every year. When new AP declarations are received, an analyst may want to identify new research projects, or may want to see the updates within a research project over time. Using an AI-driven approach, semantic search methods can be used to rank entries higher, if their content is new compared to previous years. After comparing the performance of various algorithms, term frequency–inverse document frequency (TF-IDF) scores were used as novelty detector, an established method for this type of problem, that also provides a high degree of transparency. [9]

Additionally, keywords which are unique to a declaration entry can be highlighted. A first evaluation showed that, for some States, less than 30% of their R&D-related declaration are about new research, while the other entries are describing continuations of already declared research. This result shows the high potential of data science to support analysts and decrease the required effort to review newly received declarations.

Data science can also be used to review the NFC stage of declared research, which is generally submitted by States along with a description of the research. Assistance can be provided by automatically identifying potentially incorrectly labelled descriptions (for example, a project about reactor design being mistakenly categorized as "uranium enrichment").

Currently, the main IAEA tool used to assist the AP review process (the Additional Protocol System) allows research activities to be categorised and filtered by the provided NFC stage. In some areas, a more detailed categorization can help to identify new research and can also simplify consistency analysis with open source data (see section 3.3). Therefore, an AI-enabled approach can apply automated suggested tagging of a more detailed categorization scheme, which is based on the Physical Model. The same categories and machine learning model are also used to categorise S&T publications from open sources, as explained in section 3.1 and in Schneeweiss and Stojadinovic (2022). [8]

Another method undertaken in this review process is the cross-comparison between information provided by a State and States it's collaborating (usually relating to joint research projects). If done manually, a high combination of entries, including previous years, has to be checked. Under an AI-driven approach, semantic similarities are calculated using natural language processing to find unmatched research activities which should be declared when States are collaborating. The same data science approach is used to support consistency analysis. A more detailed explanation is available in section 3.3.

A first demonstration based on a prototype shows, that the combination of described tools to support analysts can increase efficiency in verifying the correctness and completeness of declared research.

3.3. AI assistance for consistency analysis between AP and OS

During the process of consistency analysis, analysts compare if safeguards relevant open source information is consistent with State declared information, including in AP declarations. For declarations of R&D activities, the main AP articles to consider are 2.a(i), 2.a(x) and 2.b(i).

A common technique to narrow down possible declaration entries for a given open source document is a full text search using different keywords. Keywords, which are characteristic for a certain research project, are used to search for corresponding entries in the State declaration.

The data science approach to assist the analyst with this task is based on a similarity-search method. It calculates similarity scores for all possible combinations and allows the analyst to prioritize entries of the State declarations. Similar to the novelty detector implemented for the computer-assisted AP review (see 3.2), the TF-IDF method was selected as the suitable method to calculate meaningful similarity-scores. [9]

The algorithm uses similar principles to those used by analysts to search inside the State declarations. If two documents share very characteristic words, which are rare in the entire corpus, the generated similarity score will be high. In contrast, frequent words have a lower influence on the similarity score. Using results of already conducted consistency analysis from previous years, the performance of the selected approach could be evaluated.

After ranking all possible entries of declarable research, the corresponding entries were in the top 10% on average, which demonstrates the high potential to increase the effectiveness and efficiency of the consistency analysis.

In addition to the prioritization of entries of State declarations, words that are unique to a description are highlighted in the description of the declarable research, as well as in the declaration. Thus, during the process of finding semantic matches between documents, the analyst can focus at first on those keywords before continuing with a deeper investigation of the declaration.

An observation during the evaluation of the applied approach was that the highlighted keywords were usually names of specific nuclear technologies, abbreviations, project codes or organizations and locations. Those are similar types of keywords an analyst would focus on to review declarations, which confirms the high degree of transparency and usefulness of the selected data science approach.

4. CONCLUSIONS AND FUTURE WORK

To overcome the current challenge of spotting safeguards relevant information in an increasing amount of unstructured data, we presented a selection of projects based on recent breakthroughs as well as long-established data science methods. Evaluating the results gives indication of the extent to which AI can have a role for Safeguards and the challenges that have to be addressed.

In order to support Safeguards Information Analysts in their assessments, various machine learning models were trained, tested and validated on two tasks: Physical Model suggested classification and a safeguards relevance predictor. The conducted Proof of concept (PoC) demonstrated that there is potential to reduce the workload of manual S&T publications selection by more than 80% while minimizing false negatives.

One time-consuming challenge in the review of the research-related articles 2.a(i), 2.a(x) and 2.b(i) of the AP is the identification of newly declared research. A novelty detector based on TF-IDF scores prioritizes entries that describe new research content. A first evaluation showed that, for some States, less than 30% of their declarations are about new research.

Data science can also be used to highlight important keywords, suggest corrections to the provided NFC stage and to predict more detailed NFC categories. A semantic similarity metric is applied to match research projects between collaborating States, and therefore assist to verify the completeness of State declarations.

The same data science approach can support analysts with the consistency analysis, which seeks to answer whether a State's declarations are consistent with safeguards relevant information collected from open sources. An evaluation of already conducted consistency analysis shows that, after ranking declaration entries for a declarable research project, the corresponding entry showed up in the first 10% on average.

The evaluation of the described use cases demonstrates that data science can support analysts in their open source information analysis, internal consistency analysis of AP declarations and consistency analysis of those two sources of safeguards relevant information. This allows analysts to focus on tasks which require more complex analysis, such as assessing possible inconsistencies and performing appropriate following up actions.

The recent progress in the field of natural language processing is characterized by impressive innovations at a rapid speed. [10] The future work regarding the presented projects involves constant assessment if the latest methods are applicable to safeguards analysis tasks. Therefore, ethical guidelines [11], transparency, and responsible use have the highest priority, and a strong collaboration with analysts and subject matter experts during all project phases is required.

Parts of the projects to support open source analysis are already integrated in the open source information system OSIS 2.0 [7], and the other depicted use cases are available as functional prototypes. Together with analysts and subject matter experts, the evaluation of those data science projects is ongoing with a view to integrate them in software systems and analysis workflows.

REFERENCES

[1] M. Javed, A review on document image analysis techniques directly in the compressed domain, Artificial Intelligence Review: Springer, 2017.

[2] L. Bornmann, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, Humanities and Social Sciences Communications volume: Nature, 2021.

[3] S. Francis and K. Woan Jin, Science and Technology Information: A Reliable Asset or a Burden for Safeguards Implementation?, Vienna: Safeguards Symposium 2022, 2022.

[4] A. Md Zahangir, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, arxiv, 2018.

[5] C. Eldridge, T. Skoeld and G. Dupuy, Open Source Information: challenges and perspectives, Vienna: 2022 Symposium on International Safeguards, 2022.

[6] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv, 2018.

[7] T. Skoeld, F. Courbon and K. Spence, OSIS 2.0: Optimizing analyst-driven automation of open source information collection and processing for safeguards state evaluation, Vienna: 2018 Symposium on International Safeguards, 2018.

[8] P. Schneeweiss and T. Stojadinovic, Extracting the signal from the noise to support the analysis of safeguards relevant information: can artificial intelligence have a role?, Vienna: Safeguards Symposium 2022, 2022.

[9] C. U. Press, “Scoring, term weighting and the vector space model,” 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/scoring-term-weighting-and-the-vector-space-model-1.html. [Accessed 3 4 2023].

[10] A. Lancaster, “Beyond Chatbots: The Rise Of Large Language Models,” Forbes, 2023. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2023/03/20/beyond-chatbots-the-rise-of-large-language-models/.

[11] UNESCO, “Principles for the Ethical Use of Artificial Intelligence in the United Nations System,” 2022. [Online]. Available: https://unsceb.org/principles-ethical-use-artificial-intelligence-united-nations-system. [Accessed 4 4 2023].

[12] “Image Classification on ImageNet,” 28 June 2022. [Online]. Available: https://paperswithcode.com/sota/image-classification-on-imagenet.

[13] E. Gibney, “Nature,” 31 7 2022. [Online]. Available: https://www.nature.com/articles/d41586-022-01705-z.

[14] V. Sanh, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arxiv, 2019.

[15] “Evaluation of ranked retrieval results,” [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html.

[16] “Advanced Evaluation Metrics,” [Online]. Available: https://cs230.stanford.edu/section/8/.