TOSSD

Total Official Support for Sustainable Development

Methodology of the SDG artificial intelligence tool and main findings from pilot exercise on 2021 data

TOSSD Task Force Issues Paper[1] - Agenda item 3
20th meeting of the International TOSSD Task Force
7 - 9 March, 2023 - Dakar, Senegal

I. Background

1. The 2015 Addis Ababa Action Agenda and the introduction of the Sustainable Development Goals set a new standard for understanding and reporting on development flows. The 17 SDGs and their 169 underlying targets established a new conceptual framework for sustainable development, to be used by all countries to identify and subsequently direct their efforts towards the areas that are most in need.

2. Since then, discussion in the development landscape has been centred around the advancement of the SDGs, and it is common that developmental activities are conceived as contributions to the SDGs. In TOSSD, the SDG focus is an eligibility criterion, thus of utmost importance. However, determining SDG targets for TOSSD activities is a complicated task. It requires a high-level conceptual analysis, making it a time-consuming endeavour, especially when reports include thousands of activities.

3. It is in this context that the Secretariat has developed an artificial intelligence (AI) tool – the *SDG target classifier* – to provide a consistent and efficient method for assessing the SDG alignment of development co-operation activities, facilitating data reporting and verification, but also uncovering new insights in the data through the SDG lens.

II. Methodology

4. The algorithm is based on machine learning, using techniques such as *Language Models* (LMs) for semantic understanding of the reported activities and *multilabel classification* for determining the most relevant SDG targets. A warranted user may then review and compare the proposals put forth by the algorithm and adjust or reject them according to his/her interpretation of the data, therefore preserving the official nature of the reporting.

5. *The SDG target classifier* is a machine learning algorithm, meaning it is based on the concept of "learning from past examples". It exploits textual information as input to obtain a fine-grained pre-selection of the most relevant SDG targets for a given TOSSD activity. The data used for the training procedure are the TOSSD submissions that include project descriptions and project

[1] Drafted by Adriano DEL GALLO - contact: Adriano.DELGALLO@oecd.org

titles, but also sector codes, channel codes and of course SDG focus information when provided at target level. By mimicking the decision process of a human reporter, the model will gather a general sense of the (implicit) rules reporters use to assign SDG targets. The advantage is that these rules can later be applied automatically to new data entries.

6. The next three paragraphs illustrate some more advanced technical elements for data experts to consider and for those who are curious to learn more.

7. The first component of the algorithm is the *TOSSD Language model (TOSSD LM)*. This system can understand the complexity of language, the subtleties of context as well as the specificities of development co-operation vocabulary. This component is essential because it serves to understand the content of a project description. On the theoretical side of things, it achieves semantic understanding by representing words and sentences in a sophisticatedly organised mathematical space, called an *embedding* space. In this space, each data point gets assigned a position representing its meaning, such that we are able to encapsulate the meaning behind a text within a highly expressive vector. The position of this vector within the *embedding space* will determine which SDG targets the activity relates to during the classification process. The current *TOSSD LM* is a pre-trained *xlm-roberta*[2] which was domain adapted using the self-supervised Masked Language Modelling task on 700k+ project descriptions from TOSSD and the CRS and can understand activities in three languages: English, French and Spanish.

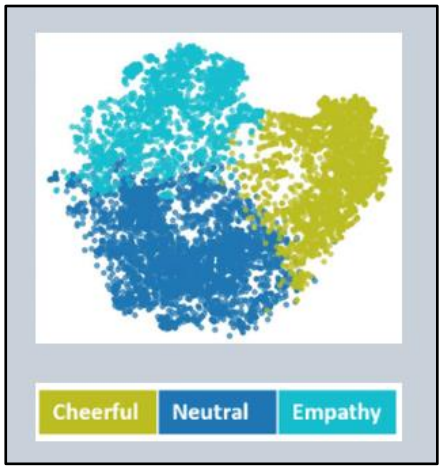


Figure 1a: Example: a 2D Visualisation of a LM's embedding space (emotion recognition task)

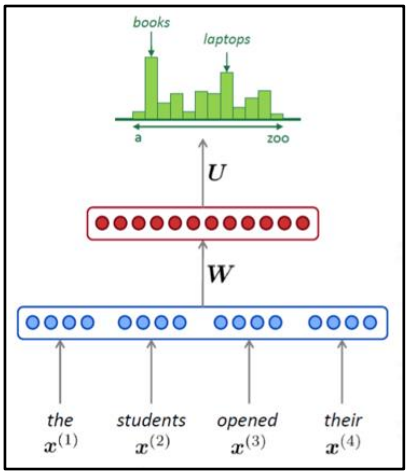


Figure 1b: Language models understand meaning

8. The TOSSD LM gives us a *mathematical representation* of the data that captures meaning. We add *a classification layer* to our Language Model for the purpose of separating this representation into distinct classes. This procedure is called the fine-tuning process: re-arranging the *embedding space* according to the SDG targets. The classification process consists of finding

[2] https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/

optimal frontiers between SDG targets in this space and assigning SDG targets based on location and boundaries.

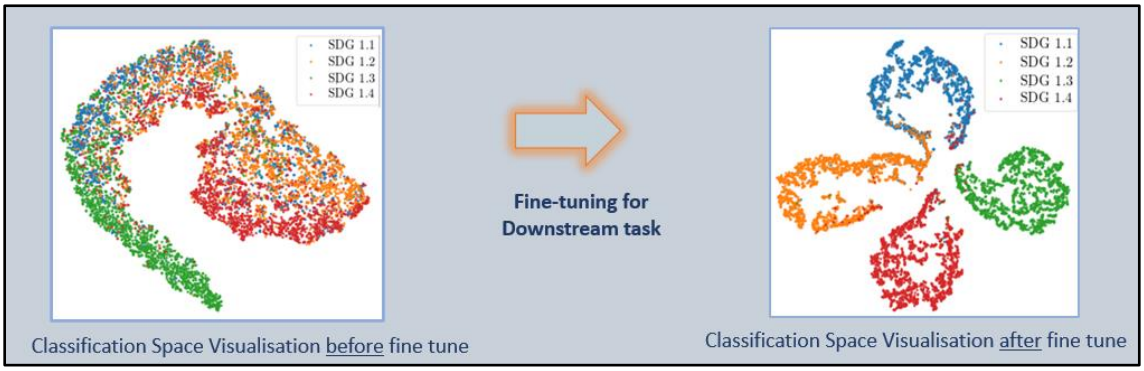


Figure 2: Fine tuning illustrated - From semantic understanding to separating SDGs

9. Regarding implementation details, the *SDG target classifier* was trained on *supervised classification* with 110k TOSSD activities that had been reported with at least one SDG at target level. We performed *data augmentation* by random contextual substitutions, to alleviate some of the class imbalances and to make the model more robust to noisy data. We also developed an *Explainability Model* based on the *Integrated Gradients* scheme to examine the relationships between the text input, the model and the predicted SDG targets. The entire architecture was coded in python using *PyTorch*[3] and the *transformers*[4] library. Data augmentation and explainability were explored using *NLPAug*[5] and *Captum*[6] .The training lasted about 15 hours on an NVIDIA A100 40GB GPU system, using *ADAM* optimiser, the *BCEwithlogit* and sigmoid for the multilabel loss, a 10e-4 *learning rate* with *linear scheduler*. The validation criterion used to select optimal model is the weighted *f-1 score* over all classes on a 20k activity validation set, and optimal probability thresholds were determined by maximising the *f-1 score* for each of the SDG targets on the same validation set.

III. Proposed use for TOSSD data

10. The *SDG target classifier* can have multiple applications for TOSSD. It can serve for data verification for reporters and the Secretariat but also for data suggestion when the *SDG Focus* needs to be filled. A dedicated website to access the algorithm is being developed – with descriptions of individual projects or excel files in the TOSSD format as an input and automated SDG target suggestions for TOSSD reporting as an output.

11. The use of the tool for data suggestion can be of interest for reporters with capacity constraints. In such cases, the tool will propose targets for activities based on the set of rules it has learnt from analysing the TOSSD database from previous years and for all reporters. These suggested

[3] https://pytorch.org/

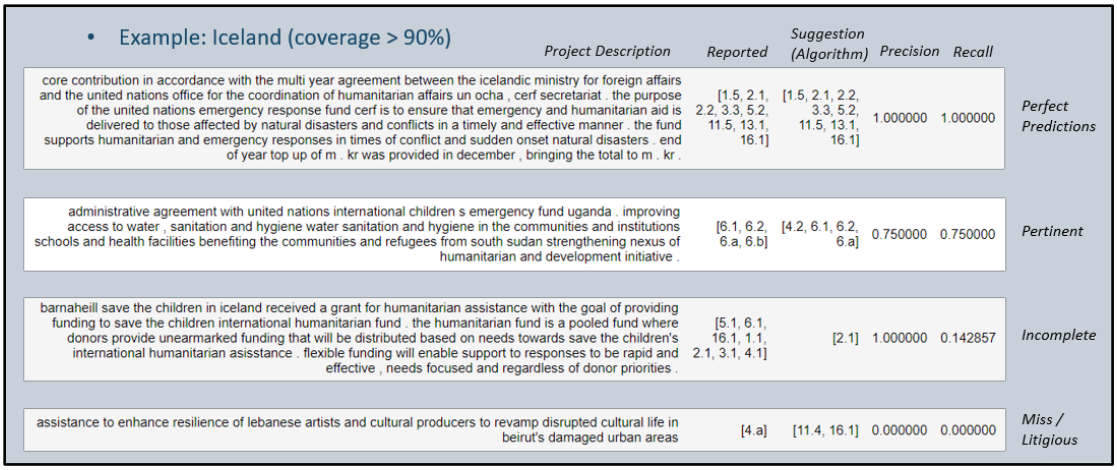
[4] https://huggingface.co/docs/transformers/index

[5] https://github.com/makcedward/nlpaug

[6] https://captum.ai/docs/introduction

SDGs, however, should be checked for coherence by a human reporter as to ensure they match with the project descriptions and the true purposes of a project. In the online tool, this aspect will be emphasised, and user will be invited to carefully consider and validate the results.

12. When applying the tool for data verification, the process remains the same, however the user will have access to additional metrics that measure the differences between reported SDGs and the AI suggestions. These metrics can be used to quickly identify possible discrepancies in the data that should be further examined or to confirm the sound choice of SDGs when both the human and AI suggestions match. These metrics are *precision* and *recall*, which respectively measure how the model's predictions are indeed part of the set of reported SDGs and how well the model can detect all the SDGs selected by reporters. The *f-1 score* is the (harmonic) mean between these two metrics which will be high when both *precision* and *recall* are simultaneously high. Indeed, a good prediction means having both a high *precision* and a high *recall*. When for a given activity, the machine's predictions are the same as the reported SDG targets, the *f-1 score* is equal to a 100% and when nothing matches the *f-1 score* will be equal to 0%.



- Example: Iceland (coverage > 90%)

Project Description	Reported	Suggestion (Algorithm)	Precision	Recall	
core contribution in accordance with the multi year agreement between the icelandic ministry for foreign affairs and the united nations office for the coordination of humanitarian affairs un ocha , cerf secretariat . the purpose of the united nations emergency response fund cerf is to ensure that emergency and humanitarian aid is delivered to those affected by natural disasters and conflicts in a timely and effective manner . the fund supports humanitarian and emergency responses in times of conflict and sudden onset natural disasters . end of year top up of m . kr was provided in december , bringing the total to m . kr .	[1.5, 2.1, 2.2, 3.3, 5.2, 11.5, 13.1, 16.1]	[1.5, 2.1, 2.2, 3.3, 5.2, 11.5, 13.1, 16.1]	1.000000	1.000000	*Perfect Predictions*
administrative agreement with united nations international children s emergency fund uganda . improving access to water , sanitation and hygiene water sanitation and hygiene in the communities and institutions schools and health facilities benefiting the communities and refugees from south sudan strengthening nexus of humanitarian and development initiative .	[6.1, 6.2, 6.a, 6.b]	[4.2, 6.1, 6.2, 6.a]	0.750000	0.750000	*Pertinent*
barnaheill save the children in iceland received a grant for humanitarian assistance with the goal of providing funding to save the children international humanitarian fund . the humanitarian fund is a pooled fund where donors provide unearmarked funding that will be distributed based on needs towards save the children's international humanitarian asisstance . flexible funding will enable support to responses to be rapid and effective , needs focused and regardless of donor priorities .	[5.1, 6.1, 16.1, 1.1, 2.1, 3.1, 4.1]	[2.1]	1.000000	0.142857	*Incomplete*
assistance to enhance resilience of lebanese artists and cultural producers to revamp disrupted cultural life in beirut's damaged urban areas	[4.a]	[11.4, 16.1]	0.000000	0.000000	*Miss / Litigious*

Figure 3: Some algorithm predictions and metrics on Iceland's 2021 TOSSD data

13. *The SDG target classifier* can furthermore be used to identify and filter data regarding specific aspects of the SDGs. This has direct applications for policy analysis, where we can use the AI tool's predictions on the TOSSD database to gather data on specific SDGs, thus giving further depth to policy analyses. This method is especially useful when dealing with data gaps and incomplete SDG alignment. The approach is currently being explored in the context of the *Aid for Trade* report and for *enhancing consideration of SDG 5 for tracking gender equality focused activities in TOSSD* (Item 7 of Task Force meeting).

IV. Results & findings from applying the SDG target classifier to TOSSD 2021 data

14. This section presents some key metrics to assess the overall average perfomance of the model to date as calculated during the collection of 2021 data, i.e. with data the model had never seen before. The scores were obtained by comparing the machine's predictions with the reference or 'ground truth' found in the reports.

a. ***Coverage***: 69706/88235 activities i.e. 79% (± 10%) -- 25 files considered

b. ***Precision / Recall score at goal level***: 76% (±8%) / 53% (±18%) – 21 files considered

c. ***Precision / Recall score at target level:*** 60% (±15%) / 47% (±19%) -- 17 files considered

d. ***Overall accuracy estimated by Task Force Secretariat:*** 82% (±7%) -- 10 files considered

Interpreting the performance scores

15. Coverage score is the percentage of activities the tool was able to assign an SDG target to. The tool is able to cover around 80% of TOSSD data, making it an effective assistant to reporters when processing very large quantities of data. There is a 10% variability between files which relates to varying quality of project descriptions.

16. The precision score reflects the pertinence of the model's predictions. Calculated at goal level it means that the predicted target matched the SDG goal reported. (Predicting 14.1 when the report indicates 14 or 14.2 will count as a correct answer.) As many activities are reported at goal level, this score is the most pertinent for assessing the general reasoning capacity of the model. This metric also considers that we can be satisfied if the predicted targets fall under the same goals as the reported targets. The precision score indicates here that the SDG target classifier will assign correct SDGs more than 75% of the time, meaning that its general capacity to identify and understand the most important SDGs present at reporting is highly reliable. In comparaison, random assigning (meaning if the machine had not learned anything) would give a correct prediction with probability 1/17 (6%) on average. There were 6 files that obtained a score over 85% on this metric and a common feature is that both descriptions and SDG reporting were exhaustive and of high quality.

17. The precision score at target level reflects the models ability to guess the exact SDG targets reported. (Here, predicting 14.2 when the report indicates 14.1 will count as a mistake.) Since there is a lot of variability and room for interpretation for the choice of a target within the same goal, this measure can be quite unstable for assessing performance. From this score, we see that the capacity of the model to assign the precise SDG target is also quite high (60%) and points to the fact that when the goal is correctly estimated, the SDG target classifier is able to have a nuanced understanding of the differences between targets and is also very good at identifying the most pertinent one. In comparaison, random assignment would give a correct prediction with a probability of 1/169 (0.6%) on average. However the high variability observed (±15%) when measuring this score does confirm that this measure is unstable for measuring overall performance of the tool, but does give us insight on the tool's good conceptual understanding of the SDG targets. There are 5 files for which this score was over 70% with a significant overlap between those and the 6 files that obtained high precision at goal level.

18. The recall score reflects the model's overall ability to find all pertinent SDGs (goal or target) associated with an activity. When comparing recall at goal (53%) and recall at target level (47%), we see that the SDG target classifier is able to find on average 50% of relevant SDGs in a project description. This score is very encouraging, however there are a number of reasons for us to believe that this metric will improve significantly in the years to come. Indeed the high variability for both these metrics (±18% and ±19% respectively) indicate both a high variability in the

exhaustiveness of SDG target reporting and also a high variability in the completeness and quality of project descriptions, and reporters can heavily influence both these aspects. On the one hand, improving quality of project descriptions (see Box 1 for our recommendations) can help the tool to unveil clear and complete infromation about the SDGs, and on the other hand, improving completeness of SDG reporting (perhaps with the use of the SDG target classifier), will prone the model to also make more exhaustive target assignments. Indeed, the data used to train the model was sourced from the early rounds of data collection, at a time when SDG target reporting was still experimental and not yet a streamlined process. As a consequence, reports from 2019 and 2020 did not always include very exhaustive accounts of the SDGs, which is a tendency therefore reflected in the model's way of assigning targets. The focus having been put on assigning the few most pertinent targets, the model was structurally biased towards precision and not recall. While SDG target reporting has substantially improved in 2021, the model had not yet learned to reflect the more exhaustive reporting style of the current data collection round, but this tendency will improve every year as we incorporate data of increasing quality when updating the model.

19. The overall accuracy estimated by the Secretariat, is a score that tries to include human feedback into the previous scoring method. For 10 data files, the Secretariat undertook a more detailed assessment of the accuracy of the *SDG target classifier*. The team processing the TOSSD data compared for each activity the pertinence of the machine predictions. For example, if a suggested SDG does not match the ground truth, the precision scores will consider the proposal false, when in practice, there could be reasonable grounds to include it based on the information available. This can happen when a suggestion is made corresponding to an SDG not included in the report but which could apply or when a suggestion falls under the proper SDG goal, though under a different target, but where the target choice is logical according to the reported information. Although this score only reflects the Secretariat's perception of SDG choice, the score being high indicates that human users tend to agree significantly with the algorithm. A notable result that illustrates the above, is that files that were annotated exhaustively by the Secretariat achieve the highest similarity with the machine predictions among all files received in the 2021 data collection. This gives the Secretariat reasons to trust the suggestions made by the algorithm, and that it can be used consistently and reliably in our work.

Common misassignments due to issues with the project titles/descriptions

20. We have identified trends in project descriptions which systematically impede the model's ability to produce correct estimates:

    a. Activities in the form of aggregates, unspecified

    b. Mistakes in spelling, typography

    c. Use of abbreviations or unusual acronyms

    d. Unclear descriptions

    e. Long descriptions with too many details that lead to confusion

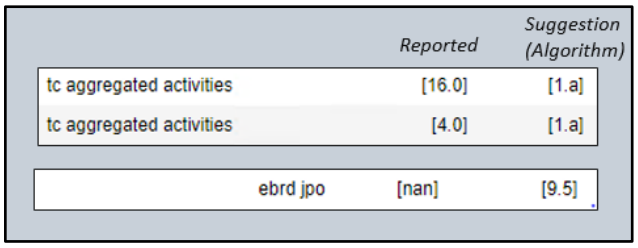


Figure 4: Brief examples of misalignment

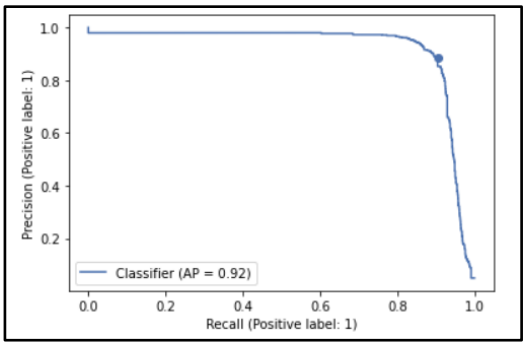
Analysing varying performance by SDG target

21. We can assess the performance of the model according to each SDG target separately, using the *area under the precision – recall curve* technique. This analysis reveals that the performance of the model can vary substantially depending on the studied SDG target (Figure 5). There are a few reasons identified which suggest a low performance:

    a. Very imbalanced overall number of activities contributing to each SDG target (Figure 6). For some targets, the model has thousands of examples to learn from, and for others there are less than a hundred examples, which is very little for the model to learn to understand and recognise the concepts.

    b. Some SDG components of activities may be omitted during reporting. Non-exhaustive SDG reporting may bias the learning process.

    c. Concepts behind targets may be general or subject to interpretation, the model therefore needs more examples to learn from (Figure 5.b)



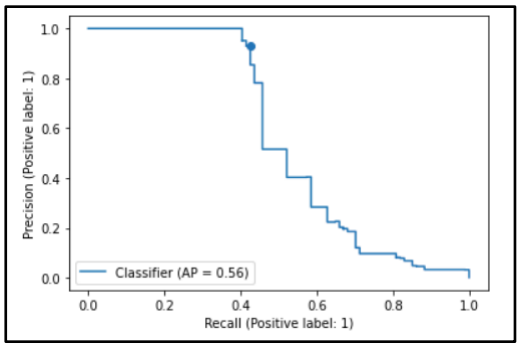


Figure 5: Comparing precision-recall curves to estimate SDG target concept understanding.
Figure 5.a (Left): SDG 3.d (6034 samples) - There exists a (blue) spot on the curve with high precision and high recall simultaneously. This implies good conceptual understanding of this SDG (f-1 = 92%)

Figure 5. b (Right): SDG 8.b (193 samples) - There is no probability threshold which allows the model to have both a high precision and a high recall simultaneously. The f-1 is 56%, but the blue spot here will be biased towards precision, meaning that when SDG 8.b is predicted it will likely be correct (around 95% precision), however it will often fail to detect it when it should (45% recall, so misses detection 55% of the time). Therefore, the model needs to progress on understanding this SDG. A direct solution would be to provide the model with more samples containing target 8.b to learn from.

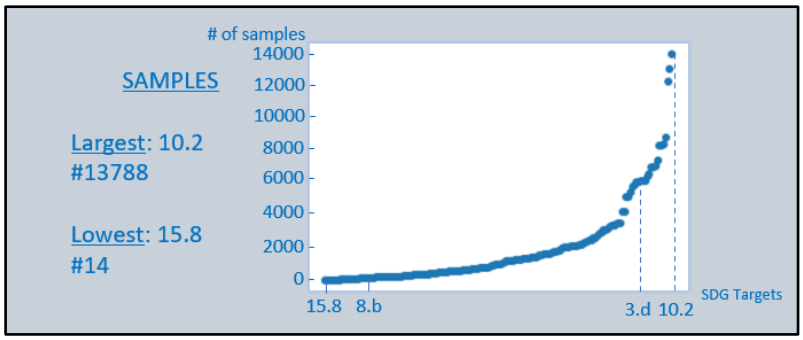


Figure 6: Number of samples in TOSSD database by SDG target - ordered by prevalence

Box 1: *Recommendations for reporting to maximise efficiency of the tool*

Quality and completeness of SDG target reporting will impact the quality of Machine Learning predictions. At every round of data collection, the model will significantly improve. The adjustments made in reporting can greatly benefit the performance of the system for the following year.

1. *Project Titles & Project Descriptions*. Detailed and concise work best (~ 1 paragraph - 3 / 4 sentences). Describing aim and context of the project rather that how resources are spent. Disaggregating some activities when possible. Being mindful of spelling, abbreviations, rare acronyms, etc.

2. When relevant, emphasise reporting at target level instead of goals, especially when activities relate to low performing SDG targets[7]. In general, providing an exhaustive or extensive list (within reason) leads to great improvements.

3. Make sure to include the necessary information used by experts to determine the targeted SDGs in the project descriptions. Sometimes the process that leads to including certain SDG targets is neither reflected in the activity description, nor elsewhere in the data we receive, leading to discrepancies between the machine predictions and the report.

[7] List of under-reported/under-performing SDG targets: 1.3, 2.5, 2.b, 2.c, 3.1, 5.4, 6.4, 6.6, 7.b, 8.1, 8.2, 8.4, 8.6, 8.7, 8.8, 8.a, 8.b, 9.2, 9.a, 9.b ,10.5, 10.6, 10.a, 10.b, 10.c, 11.a, 11.c, 12.3, 12.7, 12.a, 12.b, 12.c, 14.3, 14.6, 14.7, 14.b, 14.c, 15.4, 15.8, 15.c, 16.9, 17.12, 17.13, 17.14, 17.15, 17.19

Issues for discussion

- Does the proposed use of the *SDG target classifier* for TOSSD reporting reflect the needs of Task Force members and how can the tool be used to improve reporting? (See sections I & III.)
- Does the Task Force have comments on the methodology for proposing SDG targets using machine learning? (See sections II & IV.)
- Are members willing to reflect the suggestions made in Box 1 in their TOSSD reports to improve the *SDG target classifier*?
- Are members interested in community-of-practice type of exchanges on AI or participating in other similar exploratory work?