# Interacting Minds Centre

# Seed fund application, 2023 – Round 1

**Project name:** A scalable and explainable approach to discriminating between human and artificially-generated text

|  |
| --- |
| 1. **Applicant** (name, department and email)   Roberta Rocca  Interacting Minds Centre  [roberta.rocca@cas.au.dk](mailto:%20roberta.rocca@cas.au.dk) |
| 1. **Coworkers** (name, department/University and email) Have you collaborated previously?   Ross Deans Kristensen-McLachlan, LICS/CHC, [rdkm@cas.au.dk](mailto:rdkm@cas.au.dk)  Yuri Bizzoni, LICS/CHC, [yuri.bizzoni@cc.au.dk](mailto:yuri.bizzoni@cc.au.dk)  Rebekah Baglini, LICS/CHC, [rbkh@cc.au.dk](mailto:rbkh@cc.au.dk)  The applicants have not collaborated previously. |
| 1. **Research question**   Which linguistic and cognitive properties characterize artificially generated text? Can these features be used to build model-independent, explainable algorithms that reliably discriminate between human and artificially-generated text? |
| 1. **Approach:** (methods, sites, tools etc.)?   This project will use existing datasets of paraphrases, abstractive summarization, and dialogue generation to generate parallel datasets of human- and machine-generated texts automatically and at scale. For paraphrases, we will sample examples from the MRPC corpus – a corpus of containing human-generated paraphrase pairs – and feed sampled sentences to natural language generation (NLG) models to create artificial examples. We will adopt a similar procedure for abstractive summarization (relying on the CNN/Daily Mail corpus) and for prompt-based dialogue generation (prompts from the DailyDialog corpus) and story generation (Fan et al, 2018). This will yield a large, highly controlled dataset containing parallel human- and machine-generated examples.  We will subsequently train classifiers to discriminate between human- and machine-generated sentences using interpretable sentence descriptives as inputs. We will use interpretable linguistic and cognitive features extracted through SentSpace (Tuckute et al., 2022) and TextDescriptives (Hansen et al., 2023) as inputs to tree-based models (XGBoost) that discriminate between human vs. machine-generated texts, and inspect their SHAP values to infer feature importance. Thirdly, we will validate previous hypotheses on geometric and information-theoretic properties of machine-generated text, which suggest that model-generated texts are more similar and more repetitive than texts produced by humans. Finally, we will test human performance on the same task in an online experiment, to assess how well humans can perform text discrimination and – by inspecting relations between human performance and parametric variation of feature values – evaluate which text features are relevant for human heuristics. |
| 1. **Timeframe**   April – June 2023: Dataset generation  August – September 2023: Development and training of classification models  October – November 2023: Online experiment  December 2023: Write-up |
| 1. **Contribution to State of the Art**   Natural language generation models have become increasingly fluent, making the ability discriminate between human and artificially-generated text an urgent concern. However, existing solutions to this problem yield unsatisfactory performance, insofar as they are model-specific or are not easily interpretable for human users. This makes them unusable for any real-world applications that requireprecision and accountability, such as flagging AI-generated essays  Our project overcomes these limitations by placing a strong focus on explainability; a scalable approach to data generation; model-independence; and a comparative approach to machine-based and human discrimination. Data generated in the context of the present project will also be openly shared with the scientific community, providing benchmarks and baselines to be built upon in future research. |
| 1. **Expected outcomes**  * A high-performing, explainable text discrimination model * A paper in top NLP conferences * A large-scale dataset to be used as benchmark for text discrimination |
| 1. **Potential engagement with *interdisciplinarity***   In building an interpretable model for text discrimination based on cognitive and linguistic features, the project’s foundations combine methods and intuitions from NLP, linguistics, and cognitive science, adding an important (and, in our view, dearly needed) interdisciplinary and human-centered angle to current research on language technology.  The project team consists of researchers who already work at the intersection of these disciplines but who also come from relatively diverse academic backgrounds. This diversity of perspectives means that the project is uniquely able to incorporate both cognitive-functional approaches to discourse alongside robust experimental design involving human subjects and cutting-edge language technology. |
| 1. **Why is IMC seed funding a relevant funding instrument for the project?**   This project will pioneer a novel and highly interdisciplinary approach to a problem with increasingly high societal stakes. The seed funding scheme would provide the applicants with the opportunity to iteratively develop this innovative approach, and support future larger-scale grant application (e.g., DFF or AUFF). Furthermore, the applicants are mainly non-tenured (and newly appointed) researchers with complementary expertise and highly overlapping interests, who have never had the opportunity to collaborate before. This project would be an excellent platform to kick-start lasting and high-impact synergies between them. |
| 1. **Funding for similar projects from other sources (indicate year, funding source, total amount)**   None |
| 1. **Previous IMC Seed funding ( year, project title, outcomes)**   None |
| 1. **Additional comments**   None |
| 1. **Key references**   Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.  Hansen, L., & Enevoldsen, K. (2023). TextDescriptives: A Python package for calculating a large variety of statistics from text. *arXiv preprint arXiv:2301.02057*.  Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. *arXiv preprint arXiv:2301.10226*.  Tuckute, G., Sathe, A., Wang, M., Yoder, H., Shain, C., & Fedorenko, E. (2022). SentSpace: Large-scale benchmarking and evaluation of text using cognitively motivated lexical, syntactic, and semantic features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics* |

1. **Budget (indicate, in each row, a few words that specify the expenses)**

**How much:** Seed funding comes in two batches, up to 30.000 DKK and up to 75.000 DKK.

|  |  |
| --- | --- |
| Expenses | **2023** |
| **Salary** |  |
| Junior developers  Dataset creation and release + model training  (150hrs \* 200DKK) | 30.000 DKK |
| Student assistant  Data collection (20hrs \* 150 DKK) | 3.000 DKK |
| Payment of participants on Prolific  1000 participants \* ~1.35GBP (10 minutes) | 15.000 DKK |
| **Operational costs** |  |
| UCloud compute | 10.000DKK |
| **Total** | **58.000 DKK** |

*Main part of application (1-13) should not exceed 3 pages*

1. **Abstract suitable for dissemination at web page** (app. 1/4 page)

|  |
| --- |
| With natural language generation models becoming increasingly fluent, being able to discriminate between human and artificially-generated text has become an urgent societal problem. However, existing approaches are inaccurate, non scalable, and uninterpretable, which makes them practically unusable in real-world contexts (e.g., detection of AI-generated essays) which require precision and accountability. We propose a novel and scalable approach to training text discrimination models based on interpretable linguistic and cognitive features. Using prompts from standard NLP benchmarks for paraphrase, dialogue, and summarization, we generate parallel corpora of human- and machine-generated text, train interpretable classifiers on linguistic and cognitive descriptors, and combine insights from resulting models and experimental evidence to highlight overlaps and differences in computational and human heuristics for text discrimination. |

Send your application to Christine Parsons [Christine.parsons@clin.au.dk](mailto:Christine.parsons@clin.au.dk)

Deadline for applicants: February 17th 2023, noon.