ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

Controllable and explainable AI framework in the Automatic Assessment domain*

László Csépányi-Fürjes[a]

[a]University of Miskolc, Hungary
Institute of Information Science
laszlo.csepanyi-furjes@uni-miskolc.hu

Abstract

While significant scientific efforts have been made to develop machine learning based intelligent information systems, there was much less energy invested into studying how these systems can be made controllable and explainable. At the time of writing, we are in a situation where Artificial Intelligence (AI) and Machine Learning (ML) have become part of our everyday life as well as their negative effects, like biased responses [2] and threatening AI oppression [3]. These negative impacts on human society is seemingly being neglected by AI researchers and developers and concerned mostly by human scientists [1]. To remedy this problem this paper drafts the Synergy Unit (SU) framework, the goals of which are to make AI systems transparent and to emphasize the interests of humans in the AI model by making the system controllable and explainable. This paper also presents a prototype implementation of the proposed SU framework and shows a novel Explainable-AI solution for language-model based AI systems.

In the rare case when there is a technical proposal for the controllability issue it mostly involves implementing an external system that operates between the human user and the AI agent [3]. The problem with this kind of solution is to try controlling an AI with another AI [4]. Unlike implementing an external system this paper proposes an extension of the basic AI architecture to be able to implicitly

*This research was supported by the 020-1.1.2-PIACI-KFI-2020-00165 "ERPA - Development of RoboticProcess Automation solution for heavily overloaded customer services" project implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020 1.1.2-PIACI KFI funding scheme.

control the AI system formulating a synergy between the AI agent and the human actor.

A SU is basically a software component that is a well described and delimited set of functionalities within the AI agent, Figure 1. There are four fundamental SUs that are described in this paper, these are ExplainSU, ReportSU, ControlSU and TeachSU.

- ExplainSU: implements the ability to see and realize why a certain decision has been made by the AI system.
- ReportSU: responsible for the collection, calculation and summarization of statistical and evaluation metrics of the AI system.
- ControlSU: implements the possibility of external control.
- TeachSU: is responsible to provide extra information to the human actor to be able to learn the same skill that the AI agent has.

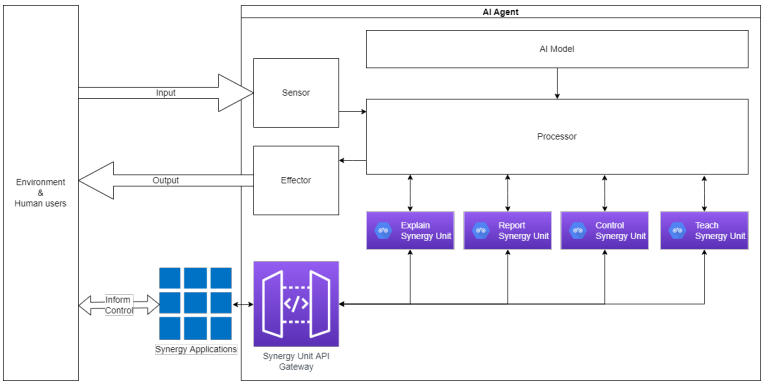

Figure 1. Synergy Unit framework.

A prototype framework was implemented as part of the Automatic Assessment (AA) module of Intelligent Tutoring Systems (ITS). The AA module evaluates textual answers by using cosine similarity $sim(X, A)$ between the expected answer $X = \{x_1, x_2, \cdots, x_m\}$ where m is the number of expected answer tokens, and the actual answer $A = \{a_1, a_2, \cdots, a_n\}$ where n is the number of actual answer tokens. For text representation the system uses the STSB-ROBERTA-LARGE language model [5]. The implemented AA module is a REST service that is responding an evaluation result. In the response the system provides a Unique Universal Identifier (UUID) as well to be able to communicate with the SU-s in another session. ReportSU provides statistics about the evaluation, while ControlSU can be used to change the evaluation schemes (TEXTUAL, BINARY, SCHOOLISH). ExplainSU uses wordcloud to put the tokens of the answer in order of relevance, Figure 2. The bigger the word in the cloud the further it takes the answer away from

the expected answer. To calculate the word relevance the algorithm skips tokens a_i of the answer one by one $A^i = \{a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_n\}$ and calculates similarity between the resulting answer and the expected answer $sim(X, A^i)$. The similarity value is in correlation with the relevance of the token.

Example question: What do the initials HAL for the HAL 9000 computer mean in the film 2001: A Space Odyssey?

Expected answer: Heuristically programmed Algorithmic computer

Actual answer: Heuristically programmed Idontknow computer

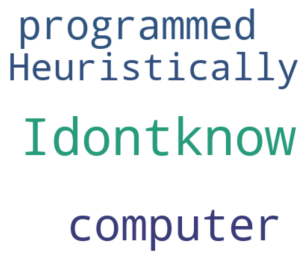

Figure 2. Explaining wordcloud

The source of the implemented prototype is available on github.

- https://github.com/csepanyifurjes/uom-su

In the future we plan to evaluate these SUs in more AI/ML applications. Also we want to collect information and experiences from end users.

References

[1] G. Csepeli: *Ember 2.0 - A mesterséges intelligencia gazdasági és társadalmi hatásai*, Kossuth Kiadó Zrt., 2020.

[2] R. Fu, Y. Huang, P. V. Singh: *Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications*, 2020, DOI: 10.1287/educ.2020.0215.

[3] G. C. Kane, A. Majchrzak: *Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants ***Forthcoming in MISQ Special Issue on Next-Generation Information Systems Theories*... The trans-nationalisation of indigenous movements: The role of digital technologies View project*, 2020, URL: https://www.researchgate.net/publication/346955741.

[4] M. Marabelli: *AI Controlling AI? A Potentially Dystopian View of Automatic Systems*, 2021, URL: https://aisnet.org/page/SeniorScholarBasket.

[5] N. Reimers, I. Gurevych: *Sentence-BERT: Sentence embeddings using siamese BERT-networks*, in: 2019, DOI: 10.18653/v1/d19-1410.