PROCEEDINGS OF SCIENCE

Exploring CrowdBots: a new evolutionary pathway for citizen science projects

Pietro Michelucci,[a,*] **Laura Onac,**[a] **Jennifer Couch,**[b] **Jacob Sherson,**[c] **Janet Rafner,**[c] **Shirley Haruka Bekins,**[a] **Roman Solovyev**[d] **and Kirill Brodt**[e]

[a] *Human Computation Institute, Ithaca, NY, USA*
[b] *National Institutes of Health, Bethesda, MD, USA*
[c] *Aarhus University, Aarhus, Denmark*
[d] *Russian Academy of Sciences, Moscow, Russia*
[e] *University of Montreal, Montreal, Canada*

E-mails: pem@hcinst.org, laura@hcinst.org

Collective hybrid intelligence may serve as an effective progression in the evolution of crowd-powered systems such as citizen science projects, which rely on human cognition. Supervised machine learning has been treated as a panacea for automated image classification under the assumption that prediction performance depends primarily on the quality and volume of the training corpus, which is often obtained through crowdsourcing. We have observed that this assumption may fall short when accurate classification depends upon contextual knowledge and abstraction capabilities. We organized a machine learning competition using data previously analysed by humans on our "Stall Catchers" citizen science platform, which gave rise to models exhibiting a range of performance characteristics. Though none of these models exhibits classification performance sufficient to replace Stall Catchers, the sensitivity and bias distributions of these models are remarkably similar to those of human volunteers, suggesting the models' suitability for crowd-based participation. We are currently conducting studies to examine the extent to which such human/AI ensembles may give purpose to imperfect ML models as an intermediate practicable step toward fully automated solutions. Our workshop sought to motivate and communicate this approach via three talks followed by a Hybrid AI simulation game, in which workshop participants broke out into teams to design hybrid AI architectures that employ CrowdBots for existing and imagined citizen science scenarios. This workshop seeded a collection of novel information processing configurations that leverage this new approach to combining AI and human cognition.

Engaging Citizen Science Conference 2022 (CitSci2022)
25-26 April 2022
Aarhus University, Denmark

[*] Speaker

PoS(CitSci2022)122

1. Introduction

Hybrid intelligence represents the collaboration of humans and machines, where artificial intelligence (AI) is augmenting human skills and creativity in order to achieve complex goals. Instead of replacing humans with AI, this technique leverages complementary strengths and learning reciprocity between humans and machines.

As hybrid intelligence may take many forms, it was the topic of a recent workshop where people that are engaged in citizen science projects in various capacities have come together to model and innovate the design of hybrid intelligence systems. The teams were provided with a set of concepts to work with and had the liberty of representing an existing hybrid process or creating a new model, for example one that introduces hybrid intelligence to an existing crowdsourcing system. This report summarizes the workshop talks, activities, and outcomes.

2. Challenges and opportunities with artificial intelligence in biomedical research

In a talk by Jennifer Couch, Branch Chief in the Division of Cancer Biology at NIH, we learned that recent technological advances have led to exciting and promising new applications of artificial intelligence in biomedical research. Still, there are a lot of challenges when applying artificial intelligence to biomedical research. For example, some AI tools used by researchers were not developed specifically for biomedical data, so their design is often missing key elements for a successful application. Another important challenge in this field is the limited interpretability of AI models. Being able to understand the decision making process of an intelligent algorithm is valuable in building trust and gaining novel biological insights.

An example of a biased intelligent algorithm was studied in [1]. They found evidence of racial bias in a commercial health management system used to assign risk levels to patients. Based on the historical data, black patients spent less money than other patients with the same level of need. Because the algorithm uses health costs as a proxy for health needs, black patients were deemed healthier than equally sick white patients.

On the flip side, interpretable machine learning models can provide valuable insights to researchers. For example, [2] developed a machine learning model able to predict the metastatic efficiency of a melanoma cell by capturing very subtle properties, impossible to identify by human experts. The interpretable nature of the model allowed for visual validation of the amplified cellular features used for prediction, revealing functional hallmarks of highly metastatic melanoma.

3. Hybrid intelligence

In a talk by Jacob Sherson, Director of the Center for Hybrid Intelligence at Aarhus University, we were presented with the current state of AI development in the industry and were introduced to the concept of hybrid intelligence. Although machine learning models can achieve state-of-the-art performance in the lab, they often fail in real world settings. According to a study performed by [3], there are many obstacles when trying to integrate AI in organizations and 70% of AI projects generate little impact. There are also risks posed by AI operating autonomously, with humans out of the loop. Bias in the training data could lead to bias amplification, relying on AI to make decisions could lead to the deskilling of humans, and training black-box systems could potentially lead to value misalignment.

As opposed to having AI be completely disruptive, in human-centered AI [4] it can be used as a tool for human augmentation. This means pursuing simultaneously high degrees of automation and also human control, with the goal of having reliable, safe and trustworthy systems. Hybrid intelligence is a variety of human computation that stipulates optimal human-AI synergies that are focused on mutual learning, with the user continuously learning from the machine, and the machine being augmented in parallel.

4. CrowdBots

Stall Catchers[1] is a citizen science project that accelerates data analysis to help understand why there is reduced brain blood flow in Alzheimer's disease patients. In mouse models of Alzheimer's disease, about 2% of the brain capillaries are stalled, causing a 30% overall reduction in brain blood flow. This is consistent with what is observed in human patients. Restoring blood flow in these mice restores memories and reduces other cognitive symptoms. [5]

Due to painstaking manual analysis required for the brain imagery in these studies, it would take between 6 and 12 months to answer one research question in the lab. Therefore, this image analysis represents a great opportunity for crowdsourcing. With over 45,000 registered participants, and 50-100 human volunteers analyzing data every day, analyzing a complete dataset can now be accomplished in only 1 or 2 months without loss of data quality.

The Stall Catchers platform allows participants to watch short clips of blood vessels in mouse brains and search for clogged capillaries where blood has stopped flowing, also known as stalls. Each player has a score and may compete on the leaderboard with the other players or teams. Playing Stall Catchers accelerates Alzheimer's disease research. Since the crowd answers exceed 99% sensitivity and 95% specificity, they are considered reliable in discriminating flowing and stalled blood vessels. On average, a reliable crowd score is obtained by combining seven separate annotations from different volunteers about the same vessel segment.

Over 45,000 Stall Catchers volunteers have contributed over 12 million classification labels for this biomedical research application. Bespoke "wisdom of the crowd" methods, which effectively create ensemble models out of humans, combined multiple individual labels for the same input to produce 1.5 million research grade labels. These gold standard data were used by over 900 participants in a machine learning competition to train 55 unique models exhibiting a range of performance characteristics. Though none of these models produces research grade labels, the sensitivity and bias distributions of these models are similar to those of individual human volunteers, suggesting the models' suitability for crowd-based participation.

An initial exploration of using machine learning models in the Stall Catchers platform involved creating a crowd bot, called GAIA, that used the third place solution from the ML competition to participate in Stall Catchers alongside human players and be a contributor to the crowd answers. A 24 hour "Catchathon" competition took place on April 28th, 2021. During this event, the crowd bot annotated at a constant rate of around 2 movies per minute and came in

[1] https://stallcatchers.com/

fourth place overall on the individual leaderboard. According to an equivalence test, the answers from the hybrid groups were more similar to the ones from an older human-only study than the answers from the new human-only groups.

A follow-up study included the prediction models from all three winners of the ML competition. So, aside from GAIA, there were two new bots participating in Stall Catchers: ZFTurbo and Clsc2. This time, instead of limiting the study to a 24-hour event, the bots and humans annotated an entire dataset that was previously analyzed only by humans. They became active on October 6th, 2021, and were stopped a bit under a month later, when each bot had analyzed all vessel segments exactly once. In this study, the performance of hybrid (ML/human) ensembles was compared to human-only ensembles. An equivalence test indicated that hybrid ensembles replicate baseline human-only data at least as well as new human-only ensembles. More analyses are underway to investigate the nuances of this finding.

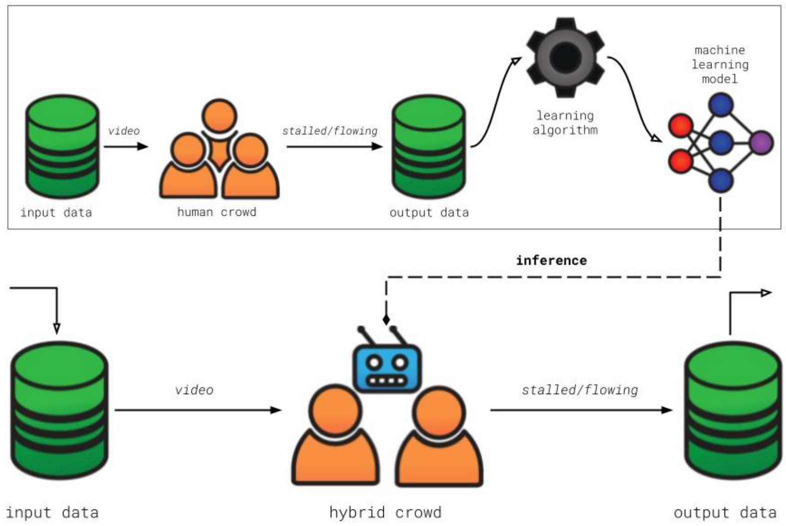

Figure 1: Stall Catchers process model.

5. Designing hybrid intelligence

In order to explore different design ideas for hybrid intelligence, we organized a workshop and formed six breakout teams tasked with creating a new process model or describing an existing one using (but not limited to) a set of predefined process elements: human, human expert, bot, human crowd, bot crowd, hybrid crowd, data, algorithm, machine learning model, and active learning. For the purpose of illustrating the use of our process elements, we prepared three use case examples of hybrid intelligence and their corresponding process models:

5.1 Stall Catchers

The CrowdBot implementation of the Stall Catchers platform is described above in the CrowdBots section. In this process, ML models, trained on human-based, crowd-generated data are then given agency as Stall Catchers participants, so they can contribute directly as members of a new kind of hybrid (machine and human) crowd, as illustrated in *Figure 1*.

5.2 iNaturalist

iNaturalist[2] is an app that helps people identify plants and animals. They use the observations to create research quality data for scientists working to better understand and protect nature. They employ a deep neural network (DNN) model to provide suggestions of species names for pictures submitted by participants, but the participants are tasked with classification of the images. In this process, AI can influence humans' responses, as depicted in *Figure 2*.

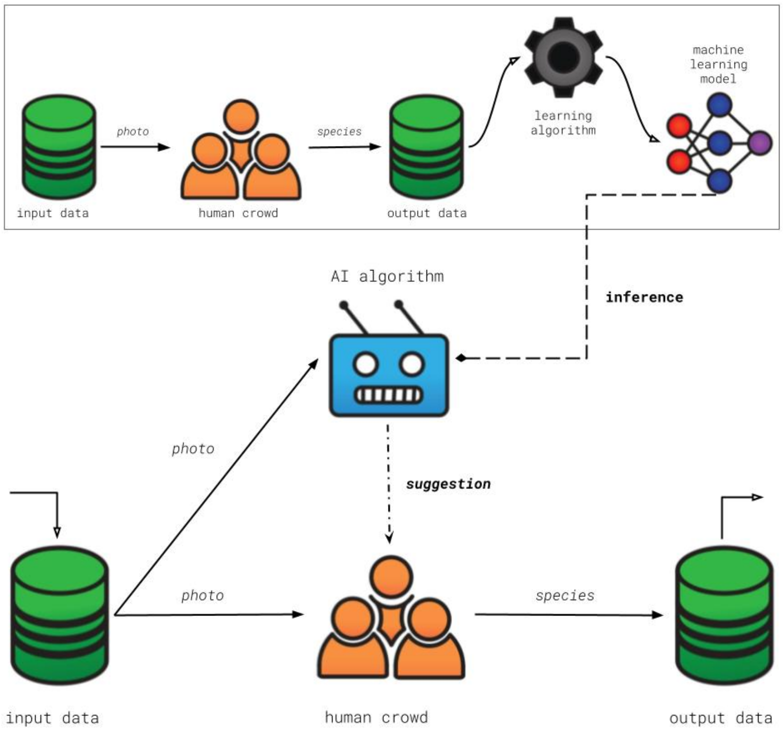

Figure 2: iNaturalist process model.

5.3 Galaxy Zoo

Galaxy Zoo[3] is an astronomy project which invites people to assist in the morphological classification of large numbers of galaxies. In Galaxy Zoo, participants classify images of galaxies according to a series of questions. Some questions are approachable with minimal domain knowledge (e.g., "Does this galaxy have spiral arms?"), while others benefit from experience (e.g., "Is there anything odd?"). The research team developed a Bayesian DNN able to learn to classify images of galaxies from volunteers' responses. They also use this model to select only the most challenging galaxies to predict for labeling by volunteers, which allows them to fine-tune models with the smallest and most informative set of labeled data. In this process, humans are annotating data in order to train the AI to solve the task and the AI decides which new data is most useful, as shown in *Figure 3*.

[2]https://inaturalist.org/
[3]https://zooniverse.org/projects/zookeeper/galaxy-zoo/

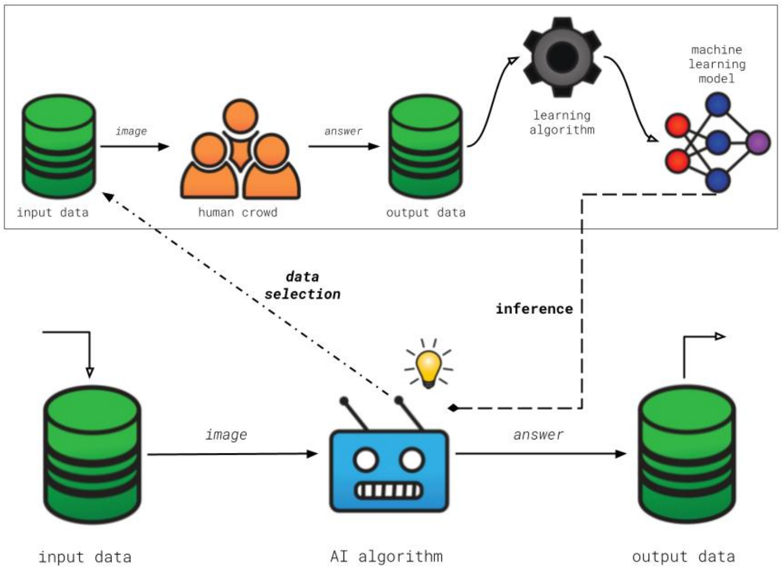

Figure 3: Galaxy Zoo process model.

6. Breakout results

The results of this exercise have been presented during the workshop and collected for further analysis.

6.1 Team 1

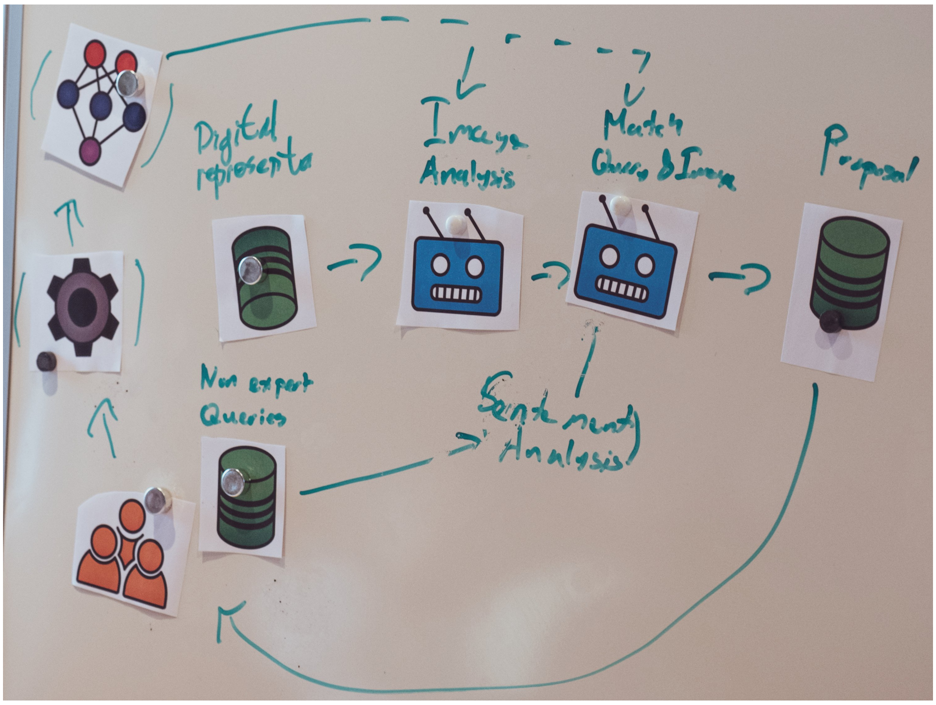

Figure 4: Team 1 process model.

Table 1 brainstormed a hybrid intelligence process, depicted in *Figure 4*, that aims to improve the customer search experience for the digital art catalog of a museum. Museum visitors (the crowd) input their non-expert search queries on the platform. For example, a visitor might input the following terms for their art interests: "happy", "family", "farm", "blue", "sunny". Each visitor's search informs an algorithm and machine learning process. This process includes data from a computer vision program that conducts image analysis of the museum art collection. The software returns images matching queries. The process includes sentiment analysis on the non-expert queries. The software then presents visitors with a list of art recommendations. Visitors indicate if their matching conditions are satisfied. This hybrid intelligence program could serve as a personalization program where the museum's AI technology learns visitors' specific art tastes and interpretations.

6.2 Team 2

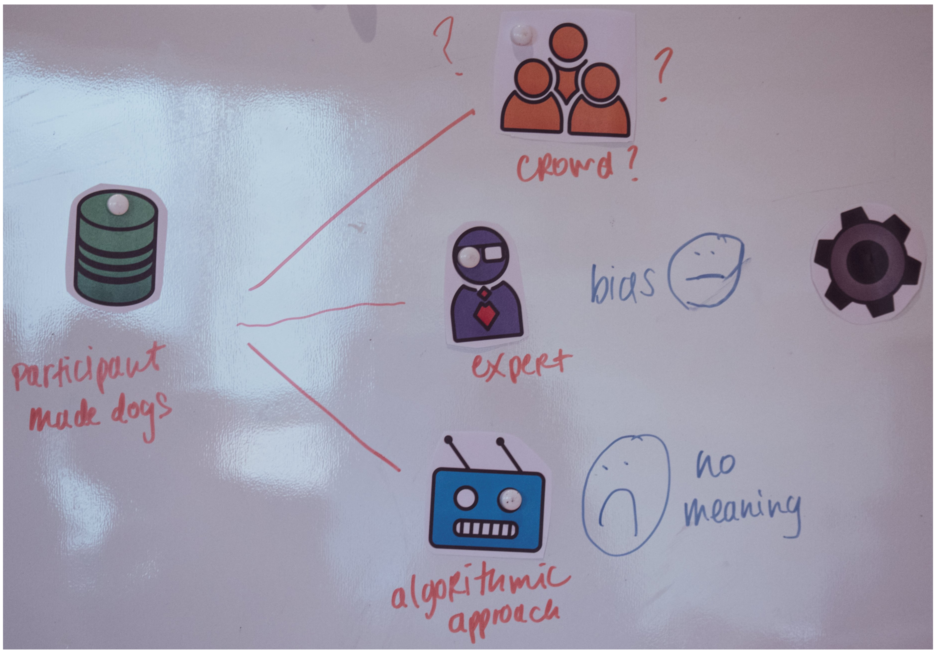

Figure 5: Team 2 process model.

Table 2 illustrated the diagram from *Figure 5* which presents a creativity project where participants are asked to create as many different dog-like figures as possible out of ten possible blocks. Afterwards, these dog depictions were evaluated based on appropriateness, flexibility, fluency, originality, and elaboration. When evaluated only by experts, there was a lot of individual bias. When the analysis was conducted by a machine using a clustering method, the results were not actually meaningful. Crowdsourcing the evaluation could result in more opinions and less bias.

6.3 Other Teams

There were four additional tables which produced creative and original process models, depicted in *Figures 6, 7, 8, 9*.

PoS(CitSci2022)122

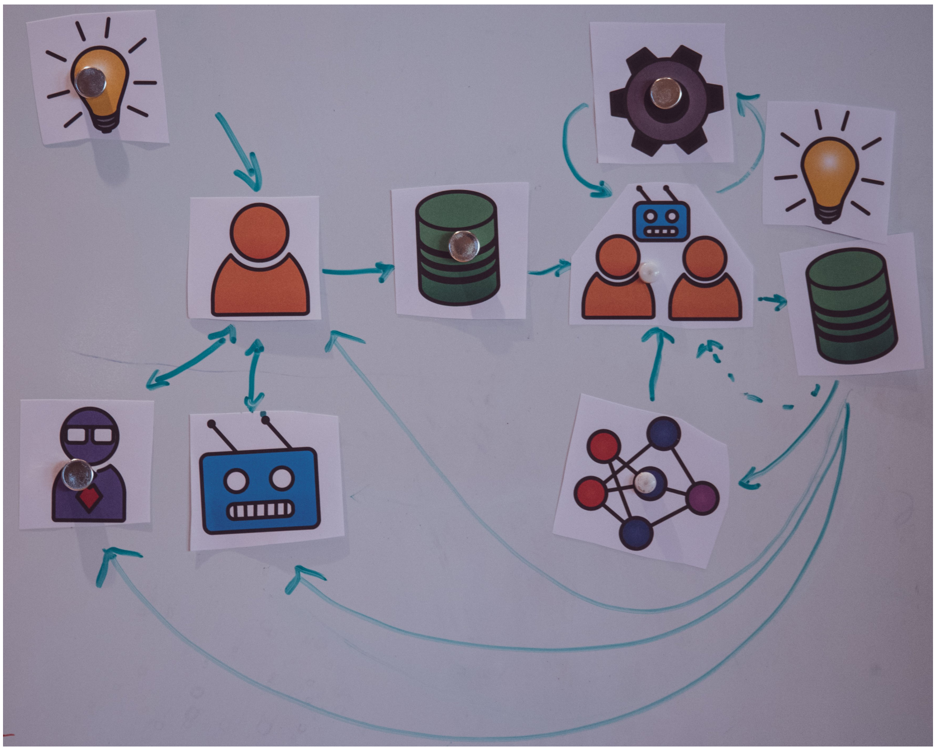

Figure 6: Team 3 process model.

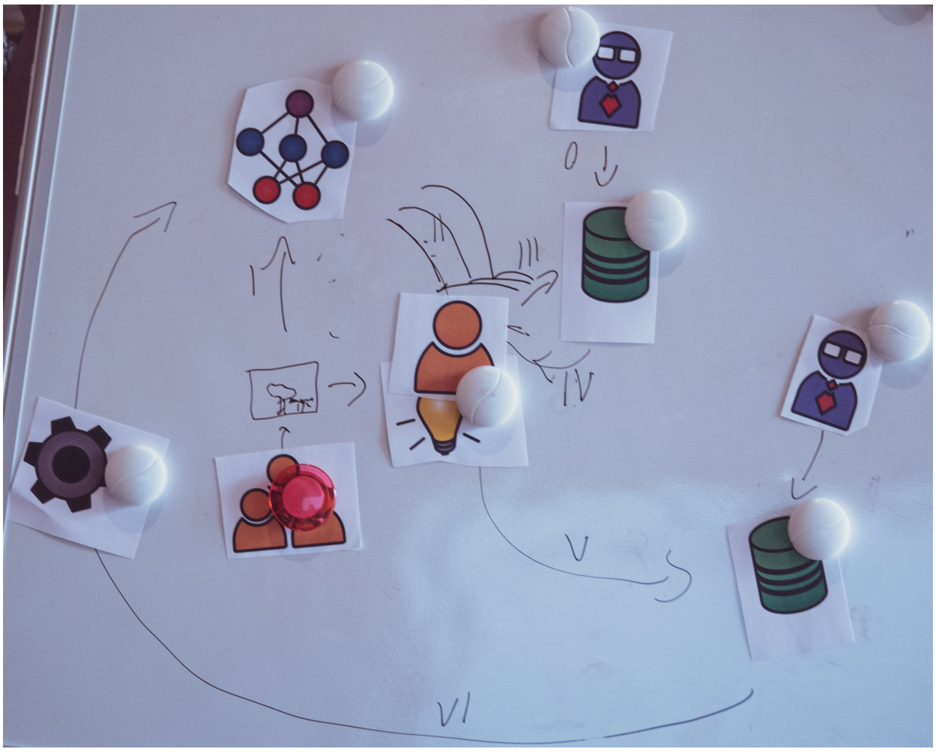

Figure 7: Team 4 process model.

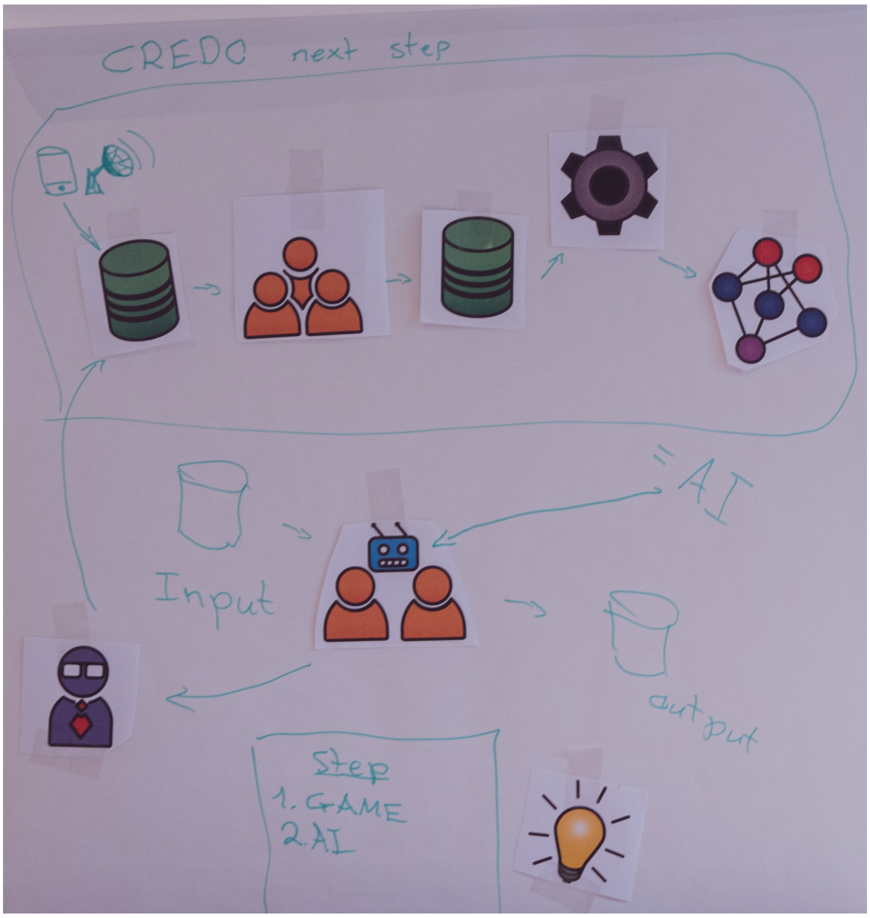

Figure 8: Team 5 process model.

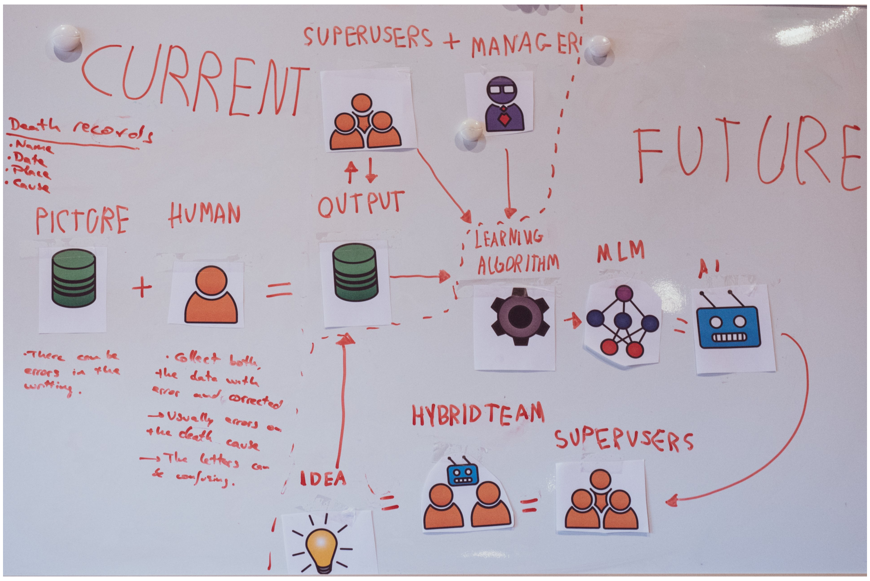

Figure 9: Team 6 process model.

7. Discussion

Recent advancements in AI continue to enable new capabilities, such as the transformer-based language models that give rise to increasingly popular machine-based envisionment platforms, such as DALL·E [6]. Despite such innovations, AI systems continue to fall short of human-like intelligence with respect to working knowledge of the world, abstraction, and creativity. Hybrid Intelligence approaches achieve futuristic AI capabilities today by seeking out information processing stages where AI struggles, and inserting human cognition into the loop.

Researchers and practitioners are accustomed to separately applying AI methods or crowdsourcing to problem-solving. Our goal with this workshop was to inspire the combined use of these approaches toward achieving new capabilities, and to orient researchers to the possibility of using such methods to tackle new issues. Several of the process diagrams generated by participants conveyed original configurations of humans and machines operating in concert toward goal-directed behavior. This outcome supports the value of tinkering with process components in the context of a conceptual framework that encourages the interplay of human and machine-based information processing. Moreover, it highlights the degree to which this space is fertile for further development with multifarious opportunities for innovation and impact.

References

[1] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* **366** (2019). https://doi.org/10.1126/science.aax2342

[2] A. Zaritsky, A. R. Jamieson, E. S. Welf, A. Nevarez, J. Cillay, U. Eskiocak, B. L. Cantarel, and G. Danuser, Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma, *Cell Systems* **12** (2021). https://doi.org/10.1016/j.cels.2021.05.003

[3] E. E. Makarius, D. Mukherjee, J. D. Fox, and A. K. Fox, Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization, *Journal of Business Research* **120** (2020). https://doi.org/10.1016/j.jbusres.2020.07.045

[4] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, *International Journal of Human–Computer Interaction* **36** (2020). https://doi.org/10.1080/10447318.2020.1741118

[5] O. Bracko, B. N. Njiru, M. Swallow, M. Ali, M. Haft-Javaherian, and C. B. Schaffer, Increasing cerebral blood flow improves cognition into late stages in Alzheimer's disease mice, *Journal of Cerebral Blood Flow & Metabolism* **40** (2020). https://doi.org/10.1177%2F0271678X19873658

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, Zero-shot text-to-image generation, *International Conference on Machine Learning* **139** (2021). https://doi.org/10.48550/arXiv.2102.12092

PoS(CitSci2022)122