Crowdsourcing Human Oversight on Image Tagging Algorithms: An initial study of image diversity

Kyriakos Kyriakou,[1,2] Pınar Barlas,[1] Styliani Kleanthous,[1,2]
Evgenia Christoforou,[1] & Jahna Otterbacher[1,2]
[1]CYENS Centre of Excellence
[2]Cyprus Center for Algorithmic Transparency, Open University of Cyprus
Nicosia, CYPRUS

Abstract

Various stakeholders have called for human oversight of algorithmic processes, as a means to mitigate the possibility for automated discrimination and other social harms. This is even more crucial in light of the *democratization of AI*, where data and algorithms, such as *Cognitive Services*, are deployed into various applications and socio-cultural contexts. Inspired by previous work proposing human-in-the-loop governance mechanisms, we run a feasibility study involving image tagging services. Specifically, we ask whether micro-task crowdsourcing can be an effective means for collecting a diverse pool of data for evaluating fairness in a hypothetical scenario of analyzing professional profile photos in a later phase. In this work-in-progress paper, we present our proposed oversight approach and framework for analyzing the diversity of the images provided. Given the subjectivity of fairness judgements, we first aimed to recruit a diverse crowd from three distinct regions. This study lays the groundwork for expanding the approach, to offer developers a means to evaluate Cognitive Services before and/or during deployment.

Introduction

Algorithms play an increasing role in society, automating aspects of our work and personal lives, taking decisions on our behalf. As people navigating a complex world, some of us develop a tendency to overtrust machines and underestimate human acumen to perform tasks (Sundar 2020). Although algorithmic processes can be critical in some situations, they can often harm individuals and groups of people in various ways. As a result, many have cited a need for monitoring algorithmic processes, where human actors constitute an important component in mitigating possible harm. The need for such monitoring arguably has become more crucial in light of the increased use of algorithmic decision making in high-risk applications, as well as in the broad dissemination and "democratization" of AI [1], e.g., through "Cognitive Services" or marketplaces such as the AI-on-Demand Platform (AI4EU).[2]

We explore the use of paid, micro-task crowdsourcing for monitoring the behaviors of vision-based cognitive services (i.e., image tagging). Rahwan (Rahwan 2018) described a vision for society-in the-loop (SITL). SITL proposes the involvement of a wide range of stakeholders, who may even have conflicting views and values, in monitoring algorithmic behaviors. When it comes to evaluating cognitive services, which often reflect social biases, involving diverse stakeholders is essential. It has been established that people approach computers with social expectations and apply human norms in evaluating computer behaviors (Nass and Moon 2000); however, social norms vary across cultures. Thus, to monitor algorithmic services, we require not only a critical mass of observers, but also significant diversity.

Our work is situated in the gap between the vision of SITL, and current practices of "human-in-the-loop" (HITL). HITL has been proposed in the machine learning and human-computer interaction communities, with the key idea being the inclusion of the human operator as a component of the system. We take inspiration from e.g., (Bansal et al. 2019; Shen and Huang 2020).

Approach

We built the OpenTag Platform as a flexible tool for understanding perceptions on the outputs of Cognitive Services (Kyriakou et al. 2020). With OpenTag, the researcher can: i) obtain informed consent from the worker; ii) decide on the prompt to instruct the participant to upload an artefact (an image in our case); iii) run the artefact through up to three Cognitive Services and return/display the results (e.g., descriptive tags); iv) configure a "survey," using four question types: *Textbox, Single Choice, Multiple Choice* or *Tag Selection* questions. Currently, we address two **research questions (RQs):** i) How we can define the diversity of the images? ii) How diverse is the set of images provided by the crowdworkers for evaluating the algorithmic service?

We asked workers to provide a *professional profile photo* to test the behaviors of Clarifai.[3] Workers then answered three questions about the way Clarifai analyzed their photo, in two similar scenarios, one higher- and one lower-risk. However, in this work, we examine only the feasibility of the approach for auditing these services via a human oversight micro-task crowdsourcing process. Specifically, we analyze the provided images in terms of their diversity.

[1]https://news.microsoft.com/features/democratizing-ai/

[2]https://www.ai4eu.eu/about-us

[3]https://www.clarifai.com/

	UK		US		IN	
Aspect	**Count**	**Ratio**	**Count**	**Ratio**	**Count**	**Ratio**
Depicting a person	80	87.91%	61	75.31%	59	67.05%
Depicting another object/entity	7	7.69%	4	4.94%	26	29.55%
Aligned with the objective	79	86.81%	61	75.31%	58	65.91%
Images likely to be themselves	15	16.48%	8	9.88%	16	18.18%
Images of other people (stock photo, celebrity)	13	14.29%	14	17.28%	11	12.50%
Invalid images (out of scope)	11	12.09%	16	19.75%	32	36.36%

Table 1: Image diversity manual analysis per location.

Worker recruitment. Aiming to get a global pool of English-speaking workers, we targeted South Africa, the Philippines, the UK, India, and the US. We found that South Africa (0 responses) and the Philippines (2 responses) did not give us enough data after waiting four days for each task to complete. In the end, we analyze data collected from India (100 responses), the US (100 responses) and the UK (88 responses). We paid $2.00/task and the median time to complete was 3 minutes and 2 seconds for UK and India, 1 minute 55 seconds for the US, and 2 minutes 45 seconds for all three regions. The tasks were run in June 2021.

Worker participation and response validity

One researcher looked over the responses at the point of HIT approval, to ensure the basic requirements had been met (from the accuracy of the survey ID the workers needed to provide, to whether responses included random key presses), and another researcher confirmed the decisions. Then, one researcher coded responses for whether they are *valid* or not. This involved checking whether images complied with the instructions on the task description. Images needed to have a professional theme; for instance, we considered images depicting a person valid, but not images depicting a fictional character or cartoon.

Diversity of images tested by workers

We coded the images based on whether they depicted a person (or another object/entity). Afterwards, for responses with images depicting people, we coded the responses for whether the image likely depicted the participant by checking their free text responses for first person language (*"I'm not looking..."*). When the participant used third person language, when the image had a stock photo watermark, and/or when the researchers recognized the person in the image as a celebrity, we coded the image as depicting other people.

We observed that a small number of participants provided an image of themselves regardless of our abstract prompts and the MTurk Use Policy[4] (16.48% of UK, 9.88% of US and 18.18% of IN contributions). This is why we included this observation as a measurable feature in our diversity framework. To identify these, we followed the same process having one researcher marking the images from each region based on the reasoning framework and then, another researcher confirmed those decisions.

[4]https://www.mturk.com/acceptable-use-policy

Diversity of the collected images. Table 1 presents our analysis. Overall, we observed that the UK and US participants complied to the instructions more closely than the India participants, who often provided images depicting different object/entities (e.g., various animals, objects like a mouse or shampoo etc.). Although the UK pool has participants who actually did our task more than once, we observe that the images provided are more *Aligned with the objective* of the task and typically *Depict a person*. In addition, UK had the least *Invalid images (out of scope)* as compared to the other regions, making the final image dataset more diverse and valid. This suggests allowing participants to make multiple contributions might actually lead to better results of higher quality in our oversight task.

Discussion and Future Work

In light of the broad dissemination of AI and AI components, it will become increasingly important to develop robust methods for auditing datasets, algorithms and services, for their social behaviors. While the human-in-the-loop approach has been used effectively for providing oversight of algorithmic processes that provide closed-ended output or judgements, evaluating processes and systems of a more open-ended nature – such as the cognitive services that analyze artefacts input by people – arguably requires the involvement of diverse stakeholders, as described in the vision for society-in-the-loop.

In our case study of a popular, vision-based Cognitive Service, the Clarifai image tagger, we demonstrated the challenges – but also the benefits – of recruiting a diverse crowd, who were asked to evaluate the service using a photo of their choice. In future work, we aim to expand our approach, envisioning a dynamic service, which would help developers, researchers and others who wish to evaluate the behaviors of AI and AI components before deploying them into an application, or even to provide ongoing oversight. To that end, in future work, we shall develop additional diversity and validity metrics, with a focus on those that could be used in a real-time, dynamic manner.

Acknowledgements

This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under agreements No. 739578 (RISE) and 810105 (CyCAT).

References

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7(1): 19. URL www.aaai.org.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2020. OpenTag: Understanding Human Perceptions of Image Tagging Algorithms. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*. Hilversum, The Netherlands. URL www.aaai.org.

Nass, C.; and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56(1): 81–103.

Rahwan, I. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20(1): 5–14. ISSN 1388-1957. doi:10.1007/s10676-017-9430-8. URL http://link.springer.com/10.1007/s10676-017-9430-8.

Shen, H.; and Huang, T.-H. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 168–172.

Sundar, S. S. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25(1): 74–88. ISSN 1083-6101. doi:10.1093/jcmc/zmz026.