A new Workflow for Human-AI Collaboration in Citizen Science

Conference Paper · September 2022

DOI: 10.1145/3524458.3547243

CITATION
1

READS
43

6 authors, including:

Ya'akov Gal
Ben-Gurion University of the Negev
81 PUBLICATIONS **1,453** CITATIONS
SEE PROFILE

Avi Segal
Hebrew University of Jerusalem
32 PUBLICATIONS **250** CITATIONS
SEE PROFILE

Eric Horvitz
Microsoft
511 PUBLICATIONS **30,447** CITATIONS
SEE PROFILE

Mike Walmsley
The University of Manchester
27 PUBLICATIONS **240** CITATIONS
SEE PROFILE

A new Workflow for Human-AI Collaboration in Citizen Science

Avi Segal
Ben-Gurion University
Beer-Sheva, Israel

Kobi Gal
Ben-Gurion University
Beer-Sheva, Israel
University of Edinburgh
Edinburgh, U.K.

Ece Kamar
Eric Horvitz
Microsoft Research
Redmond, USA

Chris Lintott
Mike Walmsley
University of Oxford
Oxford, U.K.

ABSTRACT

The unprecedented growth of online citizen science projects provides growing opportunities for the public to participate in scientific discoveries. Nevertheless, volunteers typically make only a few contributions before exiting the system. Thus a significant challenge to such systems is increasing the capacity and efficiency of volunteers without hindering their motivation and engagement. To address this challenge, we study the role of incorporating collaborative agents in the existing workflow of a citizen science project for the purpose of increasing the capacity and efficiency of these systems, while maintaining the motivation of participants in the system. Our new enhanced workflow combines human-machine collaboration in two ways: Humans can aid the machine in solving more difficult tasks with high information value, while the machine can facilitate human engagement by generating motivational messages that emphasize different aspects of human-machine collaboration. We implemented this workflow in a study comprising thousands of volunteers in Galaxy Zoo, one of the largest citizen science projects on the web. Volunteers could choose to use the enhanced workflow or the existing workflow in which users did not receive motivational messages, and tasks were allocated to volunteers sequentially without regard to information value. We found that the volunteers working in the enhanced workflow were more productive than those volunteers who worked in the existing workflow, without incurring a loss in the quality of their contributions. Additionally, in the enhanced workflow, the type of messages used had a profound effect on volunteer performance. Our work demonstrates the importance of varying human-machine collaboration models in citizen science.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

citizen science, human computer workflow

ACM Reference Format:
Avi Segal, Kobi Gal, Ece Kamar, Eric Horvitz, Chris Lintott, and Mike Walmsley. 2018. A new Workflow for Human-AI Collaboration in Citizen Science. In *ACM International Conference on Information Technology for Social Good - GoodIT 2022 7-9 September, 2022 Limassol, Cyprus.* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Citizen science engages thousands of people in solving scientific problems online[9, 20, 25]. The tremendous growth of online citizen science projects creates a need for additional labor to complete the increasing number of projects offered to volunteers, while preserving the quality of the work done. The vast majority of participants make only a few contributions before leaving [17, 26]. Thus, keeping volunteers engaged and devising ways to support their work in citizen science are important research challenges. For example, Galaxy Zoo is one of the largest citizen science projects in the world, and is the focus of our empirical investigations. Volunteers are asked to classify celestial bodies drawn from the massive Sloan Digital Sky Survey (SDSS) (see Figure 1). Analyses of Galaxy Zoo logs have shown that the vast majority of users leave quickly and make only a few contributions [20].

In this paper we describe an AI approach for addressing the engagement challenge in citizen science using an enhanced collaborative human-machine workflow. This workflow is inspired by the idea of computers working alongside people as partners in solving problems, rather than replacing human effort [8]. The workflow incorporates collaboration in two ways. First, in combining machine learning with human expertise to solve classification tasks. Machine agents uses deep neural networks to classify galaxies and turn to humans to classify "hard" galaxies, those with low classification certainty, while making mostly automatic decisions on their own for the "easy" tasks, with high classification certainty. Second, agents generate motivational messages for keeping volunteers engaged and productive that emphasize collaborative aspects of citizen science. Other works have demonstrated the benefit of using machine learning for performing tasks or for adapting motivational messages in the citizen science context [6, 22] independently. In this work, we focus on their combined effect as a collaboration platform in which humans work alongside agents to solve scientific tasks.

Introducing cooperation based workflows in citizen science systems raises several risks. First, acceptance of such workflows may be low, due to humans' reluctance to give up agency and autonomy. Second, misunderstanding may occur as to the roles of humans and machines in such new workflows. Finally, human behaviour while working along side machines may change significantly, e.g. by longer dwell time per task, poorer classification quality etc. To address these risks our approach allows volunteers to self-select whether to join the collaborative workflow and generates motivational messages that emphasize the collaborative and community beneficial aspects of human-machine work in this workflow.

We perform a controlled study to evaluate the approach in Galaxy Zoo. During the study, thousands of volunteers could select whether to join the collaborative workflow or an existing workflow in which galaxies were selected sequentially, without reference to the uncertainly of the machine learning model. We traced the behavior and contributions of the volunteers over a period of a few months. For the users selecting the collaborative workflow, we randomly assigned volunteers to cohorts that received one of four intervention messages (or no message at all) at the beginning of each session. In accordance with [21], we define a user session as a sequence of user actions where consecutive actions are not separated by more than 30 minutes.

We witnessed increased engagement and productivity of humans choosing and working in the human-machine collaborative workflow when presented with motivational messages, compared to existing workflow. Additionally, we noticed an increased dwell time per task in this workflow, possibly due the human awareness of the cooperation with the machine. Finally, we found that a motivational message emphasizing the need for humans to assist with the "hard" tasks had the most profound effect on volunteers performance.

2 RELATED WORK

Our approach builds on prior work in two separate fields of research: modeling and increasing engagement in citizen science and human-machine collaborative work in citizen science.

Past research has looked at motivating users in volunteer based crowsourcing [6, 12]. We mention studies which developed approaches for describing and extending user engagement in online communities. Anderson et al. [1] used badges to steer behavior towards required goals in question-answer sites. They modeled behavioral changes that are induced by badges for the stackoverflow site. Their model showed that change in user behavior increases as the badge frontier gets closer, and was able to predict observations about the real-world behavior of user on stackoverflow. In subsequent work, Anderson et. al [2] performed a large-scale deployment of badges as incentives for engagement in a MOOC, including randomized experiments in which the presentation of badges was varied across subpopulations.

Segal et al. [21] studied different intervention messages on the volunteers of Galaxy Zoo when the messages were chosen based on a reinforcement learning based approach. Their algorithm combined model-based reinforcement learning with off-line policy evaluation to generate intervention policies across messages and time intervals. A controlled study showed the efficacy of their proposed method and was able to outperform the state-of-the-art intervention policy for this domain, while significantly increasing the contributions of thousands of users.

Other relevant efforts come from the literature on interruption management. Horvitz et al. [10] presented a decision-theoretic approach to balancing the cost of interruptions with the cost of delay in the transmission of notifications. Shrot et al. [23, 24] used collaborative filtering to predict the cost of interruption by exploiting the similarities between users and used this model to guide an interruption management algorithm. Rosenfeld and Kraus [19] motivated and persuaded users in argumentative dialog settings using a POMDP based model and machine learning based predictions. Azaria et al. [4] considered the problem of automatic reward determination for optimizing crowd system goals and presented two algorithms that outperformed strategies developed by human experts. None of these works considered humans working alongside machines while jointly solving tasks.

Human - Machine collaborative work combines the information processing capabilities of both humans and machines [5, 14]. Many advanced approaches in this area focus on enhancing AI systems' capabilities by querying humans for feedback about a certain selection of the AI predictions [13, 16]. This becomes crucial in high stake applications, such as medical diagnostics [28].

In the Citizen Science domain, the use of Supervised Machine Learning with Deep Neural Networks (DNNs) for image classification is prominent. Here participants are tasked with labeling observations which can then be used for training better AI models, automating the task or parts of it. For example, Galaxy Zoo recently added a Bayesian DNN able to learn from volunteers to classify images of galaxies [27]. Similarly, in iNaturalist [3], a DNN model trained on scientific grade data can provide good suggestions of species names or broader taxons such as genera or families for pictures submitted by participants. The participants are also tasked with classification of the images, thus humans and AI agents are solving the citizen science tasks collectively. In [7], the authors surveyed different cooperation methods between humans and machines on joint citizen science workflows. Specifically, in a Combined workflow [29], humans and machines have the same weight when making final decisions per classification. In a Queue based workflow [30], machine learning is used as an initial step to route images to "beginner", "intermediate" or "advanced" workflows. Volunteers are allocated to these different workflows based on their estimated expertise and reliability levels. In Active Learning based workflow [27], tasks that are most informative to the machine model are selected for human classification. The efficacy of these approaches were demonstrated in offline simulations, leading to a factor of 8 increase in productivity. None of these works tested the use of motivational messages to enhance human performance while working alongside machines.

3 METHODOLOGY AND STUDY

Our methodology consists of designing and implementing the human-machine collaborative workflow within the Galaxy Zoo project and evaluating the workflow in a controlled study. We describe each step of this methodology in turn.

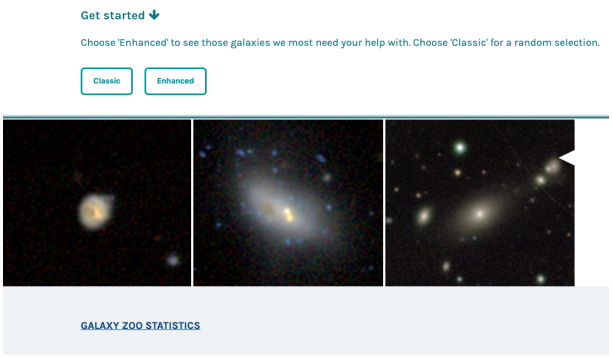

Figure 1: Option presented for selecting Classic or Enhanced workflow (partial view).

3.1 The Enhanced Collaborative Workflow

In the traditional, classic, Galaxy Zoo workflow, galaxy images for classification are allocated to users in a sequential manner, without regard to their information value, and users are not targeted with motivational messages. The collaborative workflow extends the classic workflow in two ways. First, galaxy images are automatically classified using a deep neural network [27] and assigned to a high- or low-priority queue depending on the classification uncertainty of the images. Images in the high-priority queue are more difficult to classify; they have high information value to the project and require more human classifications, while images in the low-priority queue are more straightforward and have lower information value. Galaxy images may move from the low priority to the high priority queue if human classification is ambiguous.

Second, at the beginning of each user session some of the users will receive one of four possible motivational messages that emphasize different aspects of human-machine collaboration: (1) emphasizing learning: the ability of the machine to learn from human classifications (2) emphasizing automation: the ability of the machine to automate classification of "easy" galaxies (3) emphasizing accuracy: the need for humans to review machine classifications to improve classification accuracy, and (4) emphasizing efficiency: the ability to reach more efficient classifications by combining human and machine work. New users to Galaxy Zoo are given the ability to self select one of two workflows, the classic workflow or the new collaborative workflow (see Figure 1.)

3.2 Study Design

We ran a controlled study in which this methodology was deployed in the wild in Galaxy Zoo. We compared the effects of the collaborative workflow to the classic workflow on user performance and engagement. The study received ethics approval from the Institutional Review Boards (IRB) of the University of Oxford.

The motivational messages (shown in Table 1) were developed in accordance with the administrators of Galaxy Zoo, and address different aspects of human-machine collaboration in citizen science [6]. The learning type message emphasized the value of helping to improve machine classification. The automation type message emphasized the ability of the machine to automatically classify "easy" galaxies while the human classifies the "hard" ones. The accuracy type message emphasized the need for humans to review the machine classification to improve accuracy, and the efficiency type message emphasized the efficiency that can be gained by combining human and machine work.

We made the following decisions to minimize the disruption to participants associated with the delivered messages, in accordance with guidance from the Galaxy Zoo administrators. First, we generated motivational messages *only once* per session for each user, to avoid hindering the user's work. Second, the message window included an option to opt out of receiving any additional messages. Third, the message was introduced using a window that smoothly integrated within the Galaxy Zoo GUI (see Figure 2).

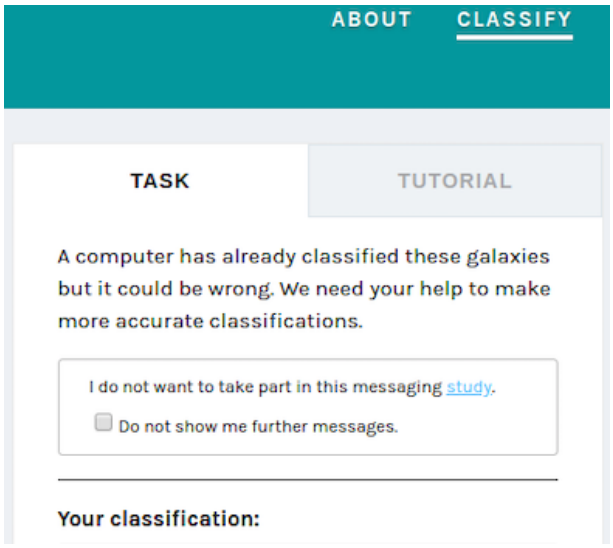

Figure 2: Intervention message in Galaxy Zoo Enhanced Workflow (partial view).

We created a cohort of enhanced workflow users for each of the four message type described above, and two additional baseline cohorts. One baseline cohort consisted of enhanced workflow users who did not receive any motivational message. An additional baseline cohort consisted of users who selected the existing classic Galaxy Zoo workflow. We hypothesized the following: (1) Users self selecting the enhanced workflow would be more productive (as measured by contributions, their quality, and users' engagement) compared to users self selecting the classic workflow. (2) The influence of the motivational messages on users' productivity depends on the type of motivational message. (3) Enhanced workflow users which were presented with motivational messages would be more productive compared to enhanced workflow users which were not presented with motivational messages.

A total of 4,971 users participated in the study. The study took place between may 23 and September 16, 2019, lasting for approximately 116 days. Users self selecting the enhanced workflow during this time period were randomly divided between the five enhanced cohorts described above (4 cohorts receiving messages and one cohort not receiving messages). Users self selecting the classic

Table 1: Intervention messages used in enhanced workflow

Type	Message	Cohort Name
Learning	A computer is learning to classify galaxies. It needs your classifications to get better!	En-Learn
Automation	We are using computer models to automatically classify easy galaxies. We need you to classify the hard ones!	En-Auto
Accuracy	A computer has already classified these galaxies but it could be wrong. We need your help to make more accurate classifications.	En-Accur
Efficiency	We are combining human and automatic computer classifications. By classifying these galaxies you are helping us to get correct answers much more efficiently.	En-Effic

workflow were allocate to the classic workflow cohorts. In total, 13,843 motivational messages were generated for all of the intervention cohorts and 273,233 classification were created by users in all cohorts during the experiment duration.

A total of 1,077 users self selected the classic workflow and were allocated to the Classic cohort. A total of 4,452 users self selected the enhanced workflow and were allocated as following: 891 users to the Enhanced-Learning and Enhanced-Automation cohorts each, and 890 users to the Enhanced-Accuracy, Enhanced-Efficiency and Enhanced-NoMsg cohorts each. Messages were presented to users at the beginning of each classification session for the cohorts that received messages and only once per session. At the request of the Galaxy Zoo administrators, we left out of the study a small minority of "super-users" with a contribution rate that was greater than three standard deviations from the mean contribution rate for all cohorts. This sub-population included 84 users (1.7% of total participants). These super users were removed from the study since they had already established themselves as persistent contributors with significantly different contribution patterns and they were not a target population for our study.

Also removed from the study were users choosing to opt out from receiving intervention messages during the experiment (456 users, 9.2%) and users which moved between the cohorts, e.g. self selecting the classic workflow during one classification session and the enhanced workflow during another classification session (560 users, 11.3%). This transitioning group will be analyzed in a separate study.

4 RESULTS

Figure 3 presents the average classification per user done in the six cohorts tested in the experiment. These include the classifications done in the classic workflow condition, in the 4 enhanced workflow conditions which received messages, and in the enhanced workflow condition which did not receive messages (En-None).

As seen in the figure, users in all enhanced workflow cohorts performed significantly more tasks per user than in the classic workflow cohort ($p < 0.05$, analysis of variance). Additionally, users in the Enhanced-Automation (En-Auto) and Enhanced-Learning (En-Learn) cohorts performed significantly more tasks on average than users in the Enhanced-None cohort (En-None) ($p < 0.05$, analysis of variance). Specifically, users in the Enhanced-Automation cohort performed on average 46.36 classifications compared to 19.48

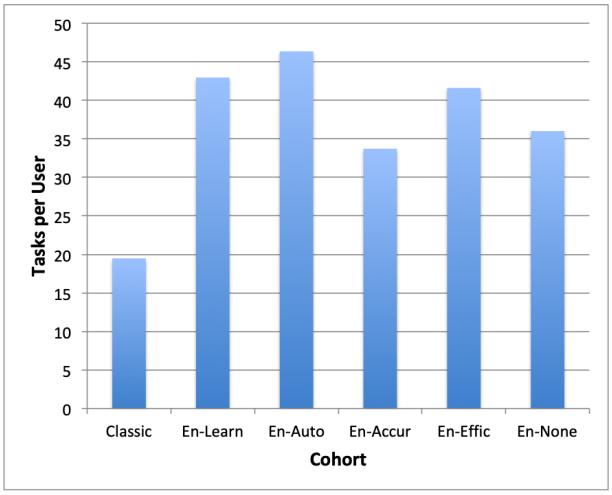

Figure 3: Comparison of classification per users for all cohorts.

classifications in the Classic cohort and 36.00 classifications in the Enhanced-None cohort.

Figure 4 compares the number of sessions performed by users in the different cohorts. As seen in the figure, all enhanced workflow cohorts users returned for more sessions than in the classic workflow cohort ($p < 0.05$, analysis of variance). Moreover, users in the En-Auto cohort returned for more sessions ($p < 0.05$, analysis of variance) than users in the En-None cohort. Specifically, users in the En-Auto cohort performed 4.12 sessions on average compare to 1.47 session and 3.08 sessions on average for the classic cohort and En-No cohort respectively. We note that there was no statistically significant difference in the number of contributions per session when comparing between the different cohorts. Thus we can conclude that the higher throughput of users in the enhanced workflow is due to returning for more classification sessions in Galaxy Zoo.

We also analyzed the dwell time (the average number of seconds between task submissions) for the different cohorts in the system (Figure 5). We found that the dwell time for all the enhanced cohorts was significantly longer than the dwell time for the classic cohort. Additionally, a statistical significant difference exists between the

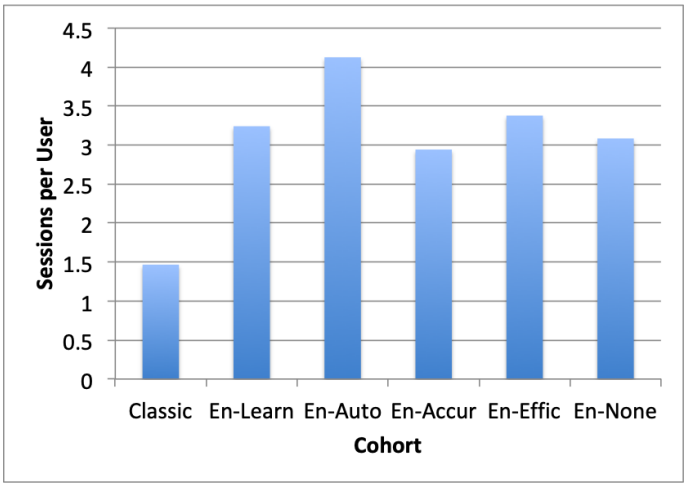

Figure 4: Average sessions per user in each cohort.

dwell time of the En-Auto and En-Accur cohorts and the En-None cohort.

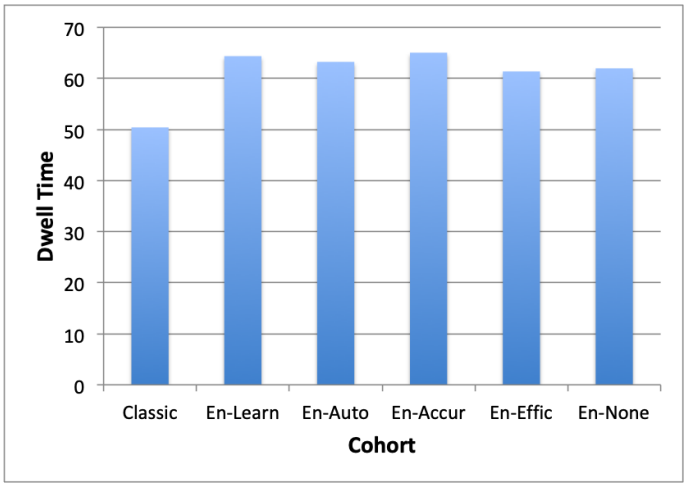

Figure 5: Average dwell time per user in each cohort.

A possible consequence of users solving more tasks in the Enhanced Workflow groups with additional dwell time is a decrease in the quality of their contributions. Since gold-standard answers to Galaxy Zoo tasks are not available, we instead used user agreement as the metric for quality. User agreement is commonly tracked as a quality metric in crowdsourcing platforms and is the basis for the aggregation algorithms such as Dawid-Skene [11]. We computed the agreement score for each cohort by iterating over all galaxies that were classified by users in this cohort. For each task, we computed the KL-divergence between the distribution of classifications collected from the different cohorts to the distribution of classifications collected from Galaxy Zoo a year prior to the beginning of our experiment and then averaged over all tasks. We found no statistically significant difference between the KL divergence of the difference cohorts. We thus concluded that the quality of the work done by users in the Enhanced Workflow cohorts was not different from that of the Classic cohort and that the increased amount of classifications did not lead to a decrease in the quality of work.

4.1 Controlling for Self-Selection

The fact that users self-selected whether to join the enhanced collaborative workflow may have biased the results in the previous section. To control for self-selection impact on our results we employ Propensity Score Matching (PSM) [18]. This method is widely used in social sciences and economics to overcome non-random assignment of treatment in evaluating social programs [15]. PSM controls for observed characteristics (covariates) related to a treatment, and measures to what extent people in either treatment or control group who share close covariates demonstrate similar results in the outcome variables. Specifically, we use PSM to compare the number of sessions performed by users in the classic cohort vs the En-Auto cohort and in the classic cohort vs the En-None cohort, while controlling for self selection. Our treatment in both analyses is the motivational message used (or omitted) and the covariates are based on users' classification behaviour during a session, including the session length, the number of tasks done in a session and the dwell time of task in a session. We note that we did not have access to demographic covariates as this information is not collected by Galaxy Zoo.

The PSM analysis is done in several steps: (1) first, logistic regression is used to compute the propensity score which is the probability of a particular volunteer choosing the enhanced workflow, given their observed covariates (2) second, volunteers are matched between the enhanced workflow and the classic workflow, based on their propensity score. Matching creates couples of volunteers, one from each group, and sampling from the majority group is used for imbalanced groups. (3) finally, the outcome of interest - in our case the average number of sessions performed by a user - is compared between the matched volunteers. Comparing the matched volunteers—one from Enhanced workflow (treatment group) and one from classic workflow (control group)—could roughly translate to having an ideal experiment, where the assignment of volunteers to workflows is random given the known covariates. Figure 6 presents the result of the PSM analysis for average number of sessions per user in the classic vs. En-None conditions (top) and classic vs En-Auto conditions (bottom). For the En-None case, we can see that after considering the Propensity Score correction, the difference between the Classic workflow and the En-None conditions in sessions per user are not statistically significant any more ($p>=0.05$, analysis of variance). Nonetheless, for the En-Auto vs. Classic analysis, the difference between the two groups remains statistically significant ($p<0.05$, analysis of variance) after the PSM correction. Thus, we note that after controlling for self selection (based on known covariates), the En-Auto intervention group continues to demonstrate a statistically significant difference of higher sessions per user compared to the classic workflow control group.

5 DISCUSSION AND FUTURE WORK

In this work, we have studied the use of a new enhanced human-machine workflow to improve productivity and engagement in the

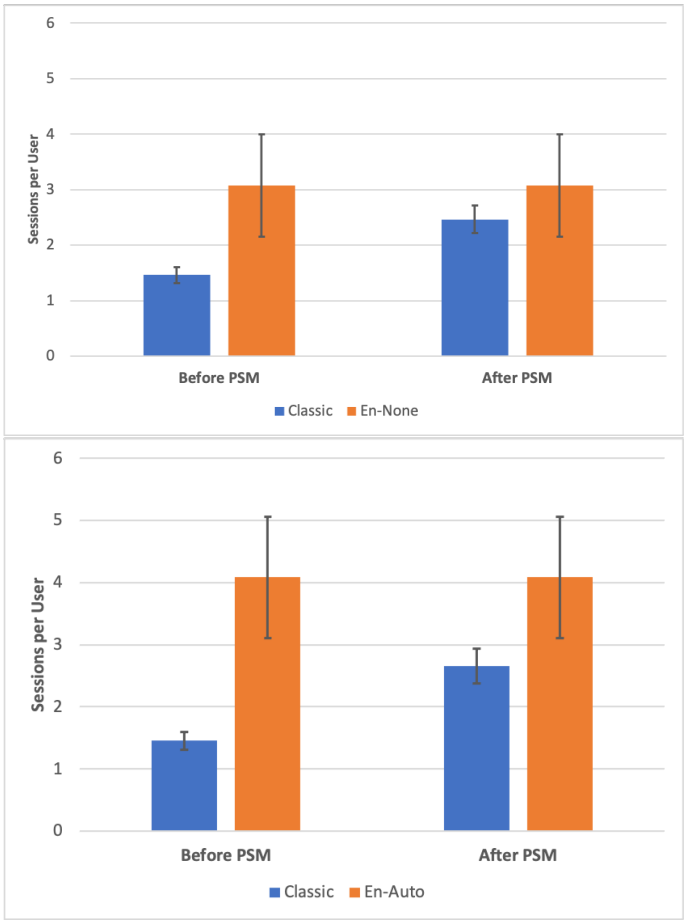

Figure 6: PSM analysis: Classic vs En-None (top) and Classic vs En-Auto (bottom)

Galaxy Zoo citizen science project. The workflow incorporates collaboration by combining machine learning with human expertise to solve classification tasks, and by using motivational messages to facilitate engagement. We evaluate this workflow in a controlled study that involved thousands of users in Galaxy Zoo. We compared the performance and engagement of users in the enhanced workflow to that of users in the existing classic workflow in Galaxy Zoo, where tasks were assigned sequentially. Our approach emphasizes autonomy and transparency, by allowing users to opt-in and opt-out at will from participating in the enhanced collaborative workflow.

We found that users self selecting the enhanced workflow return for more sessions in the system and spend more time per classification tasks compared to users self selecting the classic workflow. Additionally, we have demonstrated that the content of intervention messages is an important design choice in an intervention strategy aimed at improving volunteer engagement in the presence of an enhanced workflow. Specifically, the users presented with the Enhanced-Automation messages returned to more sessions on average in the system than users in the enhanced workflow who did not receive any message and users in the classic workflow.

One potential explanation of the influence of the automation message is that it resonates with participants' interest in making valuable contributions, which cannot be done by the machine. Specifically, this message includes a call for action ("We need you to classify") while also emphasising the uniqueness of the human expected contribution ("the hard ones!"). In contrast, is seems that messages emphasizing teaching the machine or assisting the machine to get more accurate or efficient are of somewhat lower influence. Lastly, we note the increase in dwell time across all enhanced workflow cohorts as compared to the classic cohort. We attribute this increase to the possible awareness of the human of working along side another entity which is dependant on the human quality of work.

In future work, we intend to incorporate AI in a new way into the workflow, to choose which message to generate at different points in time. To this end we need to consider the tradeoff between the immediate effect of a motivational message on users' performance, and waiting to collect more information about the user that can be used to personalize the appropriate message. We also wish to study other sets of interventions, including the modulation of task type, changing difficulty of tasks, and such factors as the specific visual properties of tasks.

6 ACKNOWLEDGEMENTS

This work was supported in part by the European Union Horizon 2020 WeNet research and innovation program under grant agreement No 823783.

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International Conference on World Wide Web*. 95–106.
[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 687–698.
[3] Maria Aristeidou, Christothea Herodotou, Heidi L Ballard, Alison N Young, Annie E Miller, Lila Higgins, and Rebecca F Johnson. 2021. Exploring the participation of young citizen scientists in scientific research: The case of iNaturalist. *Plos one* 16, 1 (2021), e0245682.
[4] Amos Azaria, Yonatan Aumann, and Sarit Kraus. 2014. Automated agents for reward determination for human work in crowdsourcing applications. *Autonomous Agents and Multi-Agent Systems* 28, 6 (2014), 934–955.
[5] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering* 61, 5 (2019), 637–643.
[6] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L Cox. 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2985–2994.
[7] Lucy Fortson, Darryl Wright, Chris Lintott, and Laura Trouille. 2018. Optimizing the Human-Machine Partnership with Zooniverse. *arXiv preprint arXiv:1809.09738* (2018).
[8] Kobi Gal and Barbara J Grosz. 2022. Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group. *Daedalus* 151, 2 (2022), 114–126.
[9] Catherine Hoffman, Caren B Cooper, Eric B Kennedy, Mahmud Farooque, and Darlene Cavalier. 2017. Scistarter 2.0: A digital platform to foster and study sustained engagement in citizen science. In *Analyzing the Role of Citizen Science in Modern Research*. IGI Global, 50–61.
[10] Eric Horvitz, Andy Jacobs, and David Hovel. 1999. Attention-sensitive alerting. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 305–313.
[11] Panagiotis G Ipeirotis. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2 (2010), 16–21.
[12] Corey Jackson, Carsten Østerlund, Kevin Crowston, Gabriel Mugar, and KD Hassman. 2014. Motivations for sustained participation in citizen science: Case studies on the role of talk. In *17th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

[13] Ece Kamar and Lydia Manikonda. 2017. Complementing the Execution of AI Systems with Human Computation.. In *AAAI Workshops*.
[14] Pietro Michelucci and Janis L Dickinson. 2016. The power of crowds. *Science* 351, 6268 (2016), 32–33.
[15] Arefeh Nasri, Carlos Carrion, Lei Zhang, and Babak Baghaei. 2020. Using propensity score matching technique to address self-selection in transit-oriented development (TOD) areas. *Transportation* 47, 1 (2020), 359–371.
[16] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 126–135.
[17] Jennifer Preece and Ben Shneiderman. 2009. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction* 1, 1 (2009), 13–32.
[18] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[19] Ariel Rosenfeld and Sarit Kraus. 2016. Strategical Argumentative Agent for Human Persuasion.. In *European Conference on Artificial Intelligence*.
[20] Avi segal, Kobi Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. 2016. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. (2016).
[21] Avi Segal, Kobi Gal, Ece Kamar, Eric Horvitz, and Grant Miller. 2018. Optimizing interventions via offline policy evaluation: Studies in citizen science. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[22] Avi Segal, Ya'akov Kobi Gal, Robert J Simpson, Victoria Victoria Homsy, Mark Hartswood, Kevin R Page, and Marina Jirotka. 2015. Improving productivity in citizen science through controlled intervention. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 331–337.
[23] Tammar Shrot, Avi Rosenfeld, Jennifer Golbeck, and Sarit Kraus. 2014. Crisp: an interruption management algorithm based on collaborative filtering. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3035–3044.
[24] Tammar Shrot, Avi Rosenfeld, and Sarit Kraus. 2009. Leveraging users for efficient interruption management in agent-user systems. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02*. IEEE Computer Society, 123–130.
[25] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282–2292.
[26] Lav R Varshney. 2012. Participation in crowd systems. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 996–1001.
[27] Mike Walmsley, Lewis Smith, Chris Lintott, Yarin Gal, Steven Bamford, Hugh Dickinson, Lucy Fortson, Sandor Kruk, Karen Masters, Claudia Scarlata, et al. 2020. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society* 491, 2 (2020), 1554–1574.
[28] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
[29] Darryl E Wright, Chris J Lintott, Stephen J Smartt, Ken W Smith, Lucy Fortson, Laura Trouille, Campbell R Allen, Melanie Beck, Mark C Bouslog, Amy Boyer, et al. 2017. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society* 472, 2 (2017), 1315–1323.
[30] Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos K Katsaggelos, Shane L Larson, et al. 2017. Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and quantum gravity* 34, 6 (2017), 064003.