SAND2019-6049C

Considerations for the Development and Implementation of Crowdsourcing to Support International Nuclear Safeguards Verification Activities

Steven Horowitz, Zoe Gastelum, Meili Swanson

Sandia National Laboratories, Albuquerque, NM, USA[*]

ABSTRACT

Active crowdsourcing - the elicitation from online crowd workers - has been successfully implemented for data collection and analysis tasks across many domains, and across sectors including academia, government, and non-governmental organizations. Several non-governmental organizations have developed nuclear non-proliferation relevant crowdsourcing activities or platforms, and the International Atomic Energy Agency (IAEA) has issued multiple “challenges” for safeguards technology development harnessing the distributed wisdom of crowds. While crowdsourcing may hold promise for multiple types of tasks that could support international nuclear safeguards verification, the construct and implementation of crowdsourcing tasks for safeguards should be carefully considered. As the culmination of over two years of research, we present a series of considerations for safeguards-focused data collection and analysis crowdsourcing tasks. Our paper considers ethical issues related to crowdsourcing in general and safeguards specifically, and practices we observed from a broad literature review to ensure the highest quality of data collection possible, including response assessments, user selection techniques, and task construction recommendations. We intend this paper to provoke thoughtful and careful considerations of task selection, deployment, and response assessment for safeguards-related crowdsourcing activities.

INTRODUCTION

Crowdsourcing is increasingly present in our daily lives. While computers are efficient modern solutions to straightforward computational analysis problems, they currently lack the ability to effectively apply reason and cognition to data. Humans are far superior at understanding and analyzing data to make inferences by applying logic and reasoning. Furthermore, the rise in machine learning has used crowdsourcing as a valuable tool for training algorithms, leveraging the relative strengths of both machines and humans. The need for large quantities of original, human-produced data for each of these cases has led to the creation of simple human intelligence tasks what can be completed in minimal time with little to no training by human users, in many cases over the Internet: crowdsourcing.

Communities across academia, government, and non-governmental and international organizations have identified and used crowdsourcing as a valuable tool to collect and assess data via tasks which are unsuitable for direct computational analysis. Fields which have leveraged crowdsourcing include scientific research and development, nature, leisure, sports, gaming, and search and rescue.

[*] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. SANDXXXXXXXX

The concept of crowdsourcing has even penetrated the nuclear nonproliferation community: The Geo4Nonpro project at the Middlebury Institute for International Studies at Monterey elicits input regarding satellite images from invited experts.[1] Crowdsourcing has also been used for community-based arms control monitoring, for example as described in an article at the Arms Control Wonk where experts used crowdsourcing to acquire supplemental data regarding Russian strategic forces.[2] Crowdsourcing has incorporated the concept of open "challenges" to promote cutting edge research and development, such as the Grand Challenges of the Defense Advanced Research Projects Agency (DARPA) which aim to solicit innovative technical solutions from a broad audience.[3] The IAEA has recently adopted a technology challenge model, holding several competitions as a method of soliciting the public's input for innovative designs and solutions to modern safeguards problems, as well as non-safeguards related issues.[4]

The recent rise in visibility of crowdsourcing raises the questions of if and how crowdsourcing could be used in support of data collection and analysis (as opposed to technology development) for international nuclear safeguards verification. Due to the high visibility of IAEA safeguards activities and the implications for states and the international community of the IAEA's safeguards conclusions, there are multiple considerations that should be carefully weighed. In this paper, we will review socio-political (including legal and ethical) considerations for potential use of crowdsourcing to collect and assess data (as opposed to conducting technology challenges) for the International Atomic Energy Agency (IAEA). Then, we will describe the data-driven considerations for the development and implementation of crowdsourcing. In doing so, this paper does not offer suggestions for direct safeguards application spaces, but rather presents a set of considerations for its mindful potential employment within this area.

SOCIO-POLITICAL CONSIDERATIONS

Socio-political considerations refer, in the context of this paper, to the legal, ethical, and political factors that a safeguards organization might consider in their use of crowdsourcing. For the purposes of our research, we cover the socio-political space by analyzing the legal background for potential use of crowdsourcing, lessons gleaned from the human studies research domain as documented in the 1979 Belmont Report, and other safeguards-specific considerations including the undue burden for member states and facility operators, the importance of independent data verification, and protection of safeguards-sensitive interests.

Legal Background

As part of the Programme 93+2, and further described in GOV/2784, the IAEA formalized its "analysis and evaluation of all relevant information available to the Agency" to support safeguards verification under its existing legal authority.[5] The use of all relevant safeguards information has been the primary means to justify the collection and analysis of open source data, including (later) social media data. However, the use of crowdsourcing to collect or assess data – that is, specifically elicit or request information that might not have otherwise existed – is an unprecedented proposition.

If the IAEA were to adopt the use of crowdsourced data or analysis, the method by which the IAEA would obtain such data would likely change the interpretation of the activity under the IAEA's legal authority. One could consider various concepts of operation under which the IAEA might have access to crowdsourced data for use in safeguards verification activities, each of which would be interpreted differently. Example concepts of operation for the IAEA obtaining crowdsourced data could include:

- The provision of crowdsourced data on a safeguards-relevant topic that a nongovernmental organization collected for its own purposes and voluntarily provided to the IAEA;
- The conception and implementation of a crowdsourcing activity by a non-Safeguards Department at the IAEA to collect geo-located imagery and environmental samples from across the world to measure environmental uptake of radioactive particles, the results of which could be shared with the Department of Safeguards once internally processed;
- The initiation of a new crowdsourcing effort by the IAEA Department of Safeguards to collect daily photographic coverage of nuclear facilities at a specific site.

Ethical Treatment of Crowdsource Workers

The 1979 Belmont Report presents basic ethical principles and guidelines under which biomedical and behavioral research should be conducted to ensure the protection of human subjects.[6] The guidelines of this report are useful in the context of crowdsourcing and establish a robust framework for the examination of the ethical issues of crowdsourcing.

The first of the ethical principles identified by the report is respect for persons, which requires individuals to be autonomous agents where their participation is voluntary, and acknowledgment and adherence to the fact that certain individuals in society possess diminished autonomy and their participation may need to be monitored, restricted, or excluded. To address respect for persons in the context of crowdsourcing, it may be prudent to rely on workers with intrinsic motivators such as belief in the nonproliferation mission and interest in the task content. Intrinsic crowd worker motivation reduces the potential for worker exploitation.

The second ethical principle of beneficence encompasses the struggle between accepting the risks associated with taking potentially progressive actions, and the possible benefits to society which may be gained by taking those risks. Safeguards crowd workers should only be asked to complete straightforward and safe tasks which do not place them in a dangerous situation physically, socially, psychologically, or economically. It should be carefully considered whether the risks are worth asking workers to conduct physically demanding tasks such as sample collection, radiation measurements, or taking outdoor photographs of facilities.

The final ethical principal described in the Belmont Report is justice. Justice means that the benefits of research should be distributed fairly, and that populations which endure the risks associated with the studies producing benefit receive at least their fair share of benefits. For monetary benefits, caution should be taken when considering distributing payments for participation and completion of crowdsourced safeguards tasks. While it is possible that users may contribute additional effort to tasks when payment is involved, task compensation may disproportionately target those in disadvantaged economic situations.

Undue Burden

With the implementation of strengthened safeguards measures resulting from Programme 93+2 activities, the IAEA recognized that the collection and analysis of additional information to support safeguards verification should not burden states with "excessive costs or by cumbersome measures to facilitate verification."[7] If the IAEA were to adopt crowdsourcing, undue burden might be realized in multiple ways. For example, if IAEA crowdsourcing for safeguards requested new images of a nuclear facility to be taken by crowd workers, those workers could potentially assemble around a sensitive perimeter, disrupting physical protection or safety operations.

Independent Verification

As with all information it collects and receives, the IAEA would independently verify the authenticity of crowdsourced information used for safeguards. Should crowdsourced data come from a third-party entity such as a non-governmental or private organization, this information would also need to be independently verified. Due to the potentially large scale of safeguards-relevant crowdsourced data, a statistical approach to information verification might lead to high confidence levels of the fidelity of the total data set. Incorrect or improperly-vetted data obtained from a safeguards crowdsourcing activity may significantly damage the reputation of the international safeguards regime.

Protection of IAEA Interests

If the IAEA were to engage in crowdsourcing for safeguards data collection and analysis, it would be conducted in a way that would not reveal sensitive information about the IAEA's interest in specific sites or area. Safeguards crowdsourcing task construction would likely be planned such that they do not reveal locations of planned or upcoming unannounced or short-notice inspections as well as not to indicate areas in which the IAEA is seeking additional information.

DATA-CENTRIC CONSIDERATIONS FOR CROWDSOURCING SAFEGUARDS

Data-centric considerations refer to measures that can be implemented in safeguards crowdsourcing activities to support the collection of usable and high-quality data for safeguards purposes (what we refer to as "quality assurance") and protecting safeguards-sensitive data. We examine quality assurance through the lens of response assessments, worker selection, and task construction. Quality assurance measures and considerations for the protection of safeguards sensitive data are below.

Quality Assurance

Various methods are available which help to raise the overall quality of data collected from crowdsourcing activities. We view quality assurance for crowdsourcing as a three-pronged approach: 1) assessing the content of the responses themselves, 2) filtering workers before granting access to a task, and 3) task construction.

Response Assessment

Response assessment refers to an assessment of the data provided by crowd workers to determine if the response should be viewed as potentially valuable information, or if it should be discarded. Response assessment allows for the filtration of spam as well as incorrect data submitted due to any number of reasons including lack of qualification of the worker, misunderstanding of the question, user error, or simply not knowing the correct response. It should be noted that analysis of the accuracy of some types of responses is difficult since many tasks may not have an objectively correct answer. For example, if users were asked to add a description/notation of an image of a centrifuge, multiple types of responses could be technically considered correct, such as centrifuge, industrial equipment, nuclear component, or enrichment.

Gold standard checking is one of the most common of the assessment techniques and includes comparing a set of known correct answers to those of the crowd.[8,9] The gold standard answers are generated either from the research team, subject matter experts, or an accepted domain standard.[10] Gold standard questions are typically placed at random between questions with an unknown response, and users are required to correctly answer these questions in order to have any of their submissions considered acceptable.[11] This typically occurs unbeknownst to the crowd worker. While this is a common method, depending on the selection and prevalence of gold standard questions, some high-quality crowd workers can fall prey to these techniques.[12]

There are other approaches to response assessment, including:

- *Majority vote* – For multiple choice responses, consideration is given to only the responses which were most selected.[13]**Error! Bookmark not defined.**
- *Manual quality checks* – This method randomly selects user responses for manual quality assurance checks by the researcher(s). This is labor-intensive and typically only occurs in smaller projects.[14]
- *Detection of nonsense answers* – Several types of nonsense answers may be submitted by users who are attempting to answer questions as fast as possible for monetary gain, or users are not paying attention. Some answers are highly unlikely given the parameters of the question or task, such as '222' for age.[15]
- *Detection of repetitive responses* – Users may attempt to deceive a system and reduce effort by copying the same language or wording from previous responses, instructions, or the task/question itself.[12,16] This spamming may be detectable through the identification of unique words relative to total words, or the 'type to token ratio.'[17]
- *Use of attention questions* – Attention questions are intended to ensure the user is properly comprehending the task instructions and/or questions. Even well-intentioned users may become bored or distracted, causing response quality to suffer. Malicious users attempting to speed through tasks can be caught with "trick questions" which require that the question is read thoroughly.[12]
- *Monitoring of task completion time* – A user's task completion time can be a useful insight into the quality of their responses. Responses given in less than a certain time threshold may be excluded[18] or sent for manual review.[19] Users which exceed a task response rate may also be excluded, depending on the task requirements.**Error! Bookmark not defined.**
- *Use of a control group* – This method gives a first group tasks to provide initial responses, and subsequently a second group is tasked with identifying their views of the validity of the responses of the first group.[13] For example, the first group may provide English-to-Spanish translations, while the second group selects if the translations are accurate or not.[20] Ideally, there would be some method of verifying that the second group was native Spanish speakers, but this will be covered under user selection.

User Selection

Rather than filtering responses, user selection can help eliminate users who are unqualified or potential spammers before they are granted access to complete a task. [21,22] Such techniques may include filtering workers by their approval rating on a platform, requiring a qualification exam, or implementing restrictions based on geographic region if the task requires specific linguistic or cultural knowledge.

Pre-filtering workers that only meet a certain worker approval rating or have successfully completed a minimum number is a common filtering strategy allowed on some platforms including Amazon's Mechanical Turk.**Error! Bookmark not defined.**[,23] Workers may be required to complete a qualification exam to:

- Test a user's ability to comprehend and follow the instructions;**Error! Bookmark not defined.**
- Qualify users for language-specific tasks;**Error! Bookmark not defined.** or
- Probe the demographic background of the user. [24], [12]

Users not meeting the metrics of the qualification exam are eliminated from task participation.[25]

Task Construction

Ideal tasks are constructed in a manner which allows for straightforward and timely completion by well-intentioned workers while simultaneously identifying and eliminating malicious workers. To discourage spammers, tasks should also ideally be designed such that malicious completion of the task takes at least as much effort as faithfully completing it.[26] Furthermore, even the most well-intentioned users can give poor responses to poorly-constructed crowdsourced tasks due to confusion, frustration, boredom, or lack of required technology. Bored users may perform more poorly than engaged users.[9]

Creating interactive tasks may help limit the number of workers who are discarded due to poor performance on attention check questions.[23] Pre-task training can demonstrate to users what to expect in the tasks and avoid confusion without having the training data count towards the actual task performance. Users may complete a pre-task assessment with gold-standard data with instant feedback regarding their responses and performance.[27] However, the training should be mindfully constructed so as to not bias the participants' responses.[27]

It is also prudent to consider how users respond when constructing the tasks. For example, while responses and data from multiple choice questions and drop-down lists may be simple to evaluate compared to open-text response, it may be more difficult to identify spoofing or spamming from malicious workers and may even attract higher rates of malicious workers if the task appears simpler. Allowing free-form or text-based responses also renders tasks vulnerable to spamming but may be easier to detect when users attempt to deceive,[12,]**Error! Bookmark not defined.** but can require additional work to consolidate the resulting data. Implementing character limits or numerical boundaries could help reduce nonsense answers.

Tasks should be constructed in a manner which minimizes necessary technological barriers. This facilitates participation by crowd workers those without access to high-quality technical equipment.[19] Furthermore, technology requirements such as high-quality microphones may result in poor-quality responses from workers not meeting assumed standards.

Finally, consideration should be given to deciding whether to compensate users for tasks. Limiting tasks to volunteers may help reduce spammers who are only interested in payment, since volunteer users may be more likely to be motivated intrinsically in the mission of the project. Compensation for completed tasks comes with multiple trade-offs, including attractiveness to spammers, perceived level of effort required by crowd workers, and the ability to recruit sufficiently large populations of crowd workers in the required time.

Protection of Sensitive Safeguards Data

The IAEA has a legal obligation to protect sensitive proprietary and safeguards information provided by a state, as well as that collected or analyzed internally (e.g., open source analysis, sampling results, and inspector observations). The collection of safeguards-relevant data may include safeguards, security, or other proprietary/privileged information. Crowdsourced safeguards data analysis activities present a risk of users establishing a mosaic or otherwise reverse engineering the project to reveal safeguards-confidential or other sensitive information.

The unintentional release of safeguard information via a crowdsourcing activity could have devastating effects, including the loss of commercial or security information from a nuclear facility which could be used for malicious purposes, and loss of trust in the IAEA which could degrade the entire system of international safeguards. It would be prudent for any crowdsourced data and activities, even in totality, be non-sensitive, such that only when combined with information internal to the IAEA (and therefore not released as part of the collection or analysis activity) would the combination become sensitive.

CONCLUSIONS

The use of crowdsourcing as a tool for the collection and/or assessment of large amounts of information has been demonstrated to be an effective strategy in multiple industries. While the IAEA has a precedent of using crowdsourcing to solicit technical innovations, crowdsourcing for safeguards data collection and analysis raises a distinct set of considerations. Crowdsourcing activities for safeguards data collection or analysis should be especially cognizant of legal and ethical implications, protecting the crowd workers from harm or abuse, and potential burden on states or sites associated with the crowdsourcing.

To support Member State buy-in for crowdsourcing for safeguards, there needs to be a demonstrated benefit. Protection of sensitive data and established information verification techniques will likely be part of the trust-building process if safeguards activities are crowdsourced.

To explore possible deployment for crowdsourcing for international safeguards, a limited form of crowdsourcing to test of its feasibility and identify additional concerns could be considered. Such a “beta” test could be released internally within the IAEA, or to trusted experts for example through Member State Support Programs.

ACKNOWLEDGEMENTS

This work was funded by the U.S. Department of Energy's National Nuclear Security Administration, Office of International Nuclear Safeguards (NA-241) Advanced Safeguards Approaches and Assessments portfolio. Special thanks to Melissa Einwechter, Cameron Stanuch, and Wayne Mei for their support of this work.

The authors also wish to thank Kari Sentz and Cristina Rinaudo from Los Alamos National Laboratory for their collaboration in this research.

END NOTES

[1] Hanham, M., Dill, C., Lewis, J., Kim, B., Schmerler, D., & and Rodgers, J. "OP#28: Geo4nonpro.org: A Geospatial Crowd-Sourcing Platform for WMD Verification." *Middlebury Institute of International Studies at Monterey*, June 30, 2017, https://www.nonproliferation.org/op28-geo4nonpro-org-a-geospatial-crowd-sourcing-platform-for-wmd-verification/.

[2] Lewis, J. "Crowdsourcing Russian ICBMs." *Arms Control Wonk*, April 2, 2014, https://www.armscontrolwonk.com/archive/207196/crowdsourcing-russian-icbms/.

[3] Defense Advanced Research Projects Agency. "Prize Challenges." https://www.darpa.mil/work-with-us/public/prizes.

[4] International Atomic Energy Agency. "Challenges." https://challenge.iaea.org/challenges/all.

[5] International Atomic Energy Agency. (1995). "Strengthening the Effectiveness and Improving the Efficiency of the Safeguards System: A Report by the Director General." GOV/2784.

[6] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Office of the Secretary. (1979). "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research." https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

[7] International Atomic Energy Agency. (1995). "Strengthening the Effectiveness and Improving the Efficiency of the Safeguards System: A Report by the Director General" GOV/2784, pp. 1-2.

[8] Smucker, M. D., & Jethani, C. P. (2011). The crowd vs. the lab: A comparison of crowd-sourced and university Laboratory participant behavior. In *Proceedings of the SIGIR 2011 Workshop on crowdsourcing for information retrieval*, Beijing, China, 28 July.

[9] Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments, *Information retrieval*, 16(2), 138-178.

[10] Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, Philadelphia, PA, 29–31 March (pp. 557-566). ACM.

[11] Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., & Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh*, Scotland, UK, 27–31 July (pp. 670-679). Association for Computational Linguistics.

[12] Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea, 18-23 April (pp. 1631-1640). ACM.

[13] Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, 57(11-12), 2918-2932.

[14] Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision, 101*(1), 184-204.

[15] de Winter, J. C. F., Kyriakidis, M., Dodou, D., & Happee, R. (2015). Using CrowdFlower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, 3, 2518-2525.

[16] Eickhoff, C., & de Vries, A. (2011). How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*, Hong Kong, China, 9-12 February (pp. 11-14).

[17] Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual English-German image descriptions. *arXiv preprint arXiv:1605.00459*.

[18] Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2015). Mechanical Turk-based experiment vs laboratory-based experiment: A case study on the comparison of semantic transparency rating data. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, 30 October-1 November (pp. 53-62).

[19] Caines, A., Bentz, C., Graham, C., Polzehl, T., & Buttery, P. (2015). Crowdsourcing a multi-lingual speech corpus: recording, transcription, and natural language processing. In *Proceedings of INTERSPEECH*.

[20] Negri, M., & Mehdad, Y. (2010). Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, 6 June (pp. 212-216). Association for Computational Linguistics.

[21] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, Perth, Australia, 3-7 April (pp. 1391-1399). International World Wide Web Conferences Steering Committee.

[22] Saupe, D., Hahn, F., Hosu, V., Zingman, I., Rana, M., & Li, S. (2016). Crowd workers proven useful: A comparative study of subjective video quality assessment. In *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*, Lisbon, Portugal, 6-8 June.

[23] Brigden, N. (2015). Interactivity and Data Quality in Computer-Based Experiments. In K. Diehl and C. Yoon (Eds.), *NA – Advances in Consumer Research* (Vol. 43, pp. 18-22). Duluth, MN: Association for Consumer Research.

[24] Li, H., Zhao, B., & Fuxman, A. (2014). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, Seoul, South Korea, 7-11 April (pp. 165-176). International World Wide Web Conference Committee (IW3C2).

[25] Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, 6 June (pp. 139-147). Association for Computational Linguistics.

[26] Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies With Mechanical Turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems*, Florence, Italy, 5-10 April (pp. 453–456). ACM.

[27] Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, Geneva, Switzerland, 19-23 July (Vol. 2126).