As the focus of ML (machine learning) has shifted toward settings characterized by massive datasets, researchers have become reliant on crowdsourcing platforms. Just for the NLP (natural language processing) task of passage-based QA (question answering), more than 15 new datasets containing at least 50k annotations have been introduced since 2016. Prior to 2016, available QA datasets contained orders of magnitude fewer examples.

The ability to construct such enormous resources derives, largely, from the liquid market for temporary labor on crowdsourcing platforms such as Amazon Mechanical Turk. These practices, however, have raised ethical concerns, including (1) low wages; (2) disparate access, benefits, and harms of developed applications; (3) reproducibility of proposed methods; and (4) potential for unfairness and...
discrimination in the resulting technologies.⁹,¹⁴

This article looks at what ethical framework should govern the interaction of ML researchers and crowdworkers, and the unique challenges in regulating ML research. Researchers typically lack expertise in human-subjects research and require guidance on how to classify the role crowdworkers play to comply with relevant ethical and regulatory requirements. Unfortunately, clear guidance is lacking: Some institutions and a 2021 paper by Shmueli, et al. suggest that all ML crowdworkers constitute human subjects;²³ others suggest that ML crowdworkers rarely constitute human subjects.¹⁰ Confusion surrounding ML crowdworkers is grounded in the following factors.

- **Novel relationships.** The U.S. Common Rule was developed in the wake of abuses in biomedical and behavioral research and reflects the need to distinguish clinical research from medical practice.¹⁵ Because the distinction between employees on a research team and study participants is less ambiguous in medical contexts, little attention has been paid to criteria for distinguishing research staff from study participants.

- **Novel methods.** In biomedical or social sciences, data is collected to answer questions that have been specified in advance, while ML often involves a dynamic workflow in which data is collected in an open-ended fashion and research questions are articulated in light of its analysis. Additionally, ML researchers often release rich data resources, where much of the data is not analyzed.

- **Ambiguity under the Common Rule.** Whether an
individual is a human subject hinges on whether the data collected, and later analyzed, is about that individual. As Shmueli, et al. have noted, crowdworkers can fill such diverse roles in ML research that it becomes difficult to draw a line between collected data about the crowdworkers versus merely from them (but about something else).  

» Scale. NLP research produces hundreds of crowdsourcing papers per year, with 703 appearing at the top venues alone from 2015-2020.

» Inexperience. Crowdsourcing-intensive ML/NLP papers seldom discuss ethical considerations that would otherwise be central to human-subjects research, and they rarely discuss whether IRB (institutional review board) approval or exemption was sought—only 14 (about 2 percent) of the aforementioned 703 papers described IRB review or exemption.

CURRENT REGULATORY FRAMEWORK
In the United States, the regulations governing the treatment of humans in scientific research, detailed in the CFR (Code of Federal Regulations), are known as the Common Rule. Falling under the auspices of OHRP (Office of Human Research Protections) of the U.S. Department of Health and Human Services, these regulations apply only to institutions that accept federal funds or have agreed to abide by these rules. Two important criteria determine whether a person constitutes a research participant: those that define research and those that define a human subject. Research is defined, in part, as “a systematic investigation, including research development, testing,
and evaluation, designed to develop or contribute to generalizable knowledge.”

A human subject is defined as “a living individual about whom an investigator (whether professional or student) conducting research: (i) obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or (ii) obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens” [45 CFR 46.102 (e)(1)].

For simplicity, this discussion is limited to the production of information, rather than to a discussion of specimens.

Two points of clarification: First, to satisfy the definition of a human subject in the CFR, researchers must retrieve data about an individual. This doesn’t imply that the study focuses on the individual but aims to generate generalizable knowledge. For example, in biomedicine, individual measurements are used to produce knowledge about a wider population. Defining what information is about an individual can be challenging for ML researchers dealing with crowdworkers.

Second, conditions [i] and [ii] in the CFR lump together a range of cases that vary in substantive ways. Condition [i] is a combination of two conjuncts. The first concerns the way that information is produced: from either intervention or interaction. These terms are defined as:

- Intervention includes both physical procedures by which information or biospecimens are gathered (e.g., venipuncture) and manipulations of the subject or the subject’s environment that are performed for research purposes.
• Interaction includes communication or interpersonal contact between investigator and subject.

Of these, interaction is the weaker condition. Interventions can be understood as the subset of interactions that produce a change in either the individual (e.g., administering a drug or drawing blood) or their environment (e.g., placing an individual in an imaging device). In contrast, interactions include communication or interpersonal contact that generates information without necessarily bringing about a change to the individual or their environment. For example, a study might divide participants into two groups: one to test an intervention alongside usual care; one to receive just the usual care. The group receiving only usual care is still part of the study’s social interaction that generates data to control for confounding, thus aiding in creating generalizable knowledge.

The second conjunct in condition (i) requires that information arising in one of these two ways—intervention or interaction—is then used, studied, or analyzed. Of these, use is the broadest category, as there may be myriad ways that information from a social interaction is used in research. In contrast, study and analysis constitute a stricter subset of uses in which data are analyzed or evaluated, presumably to generate the generalizable knowledge that defines the study in question.

Table 1 lists combinations from these categories forming different research paradigms. Among these, the intervention analysis condition is narrowest, which implies that a person becomes a study subject through targeted interventions and subsequent analysis. In contrast, the interaction use criteria are broader, holding that a person is
a human subject if, in the course of research, researchers interact with them in a way that produces information used to further the goals of the research.

Condition (ii) of the CFR’s definition of human subject applies when researchers obtain, use, study, analyze, or generate private information about a living individual, even if direct interaction is absent. It covers research involving datasets containing personal information or studies generating such information from noninclusive datasets.

These definitions demarcate which set of ethical and regulatory requirements applies to an activity. Activities not involving human subjects are not governed by regulations for human subject research, making IRB review unnecessary. Research involving human participants, however, necessitates adherence to specific moral and regulatory responsibilities, including mandatory IRB review.

This last claim might come as a surprise to some familiar with the Common Rule, since a significant portion of ML research, and NLP research in particular, is likely to be classified as exempt. Per 46.104.(3)(i) of the Common Rule, research involving benign behavioral interventions...
in conjunction with the collection of information from an adult subject through verbal or written responses or audiovisual recording can qualify for exempt status if the subject prospectively agrees to the intervention and information collection and at least one of the following criteria is met:

- The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects.

- Any disclosure of the human subjects’ responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects’ financial standing, employability, educational advancement, or reputation.

However, a researcher cannot unilaterally declare their research to be exempt from IRB review. Rather, **exempt** is a regulatory status that must be determined by an IRB (§46.109.(a)). This may seem paradoxical, as for a study to qualify for exempt status, researchers are obligated to offer comprehensive details regarding their project to the IRB. The board assesses this information to ensure all applicable Common Rule standards are met. This is common in administrative rulemaking, as well as judicial review; courts may determine whether something is in their jurisdiction, but a plaintiff has to provide information to enable a court to make that determination. Exempt status usually entails less effort and receives faster approval than a full IRB review. A researcher at an institution governed by the Common Rule would violate regulatory obligations by
commencing human subject research without prior IRB review, even if the research would have been exempt.

COMMON RULE AND ML RESEARCH
Based on the preceding analysis, there is a large subset of ML research in which crowdworkers are clearly human subjects. These cases fit squarely into the paradigm of research, familiar in biomedicine and social science, where researchers interact with crowdworkers to produce data about those individuals, and then analyze that data to produce generalized knowledge about a population from which those individuals are considered representative samples.

In some studies, researchers assign crowdworkers at random to interventions to produce data that can be analyzed to generate generalizable knowledge about best practices for using crowdworkers. Here, crowdworkers are clearly human subjects. They are the target of an intervention designed specifically to capture data about them and their performance.

For example, Khashabi, et al. engaged crowdworkers to investigate which workflows result in higher-quality QA datasets. They recruited one set of crowdworkers to write questions given a passage, while another group of crowdworkers were shown a passage along with a suggested question and were tasked with minimally editing this question to generate new questions. In these settings, the data was about the workers themselves, as was the analysis.

Similarly, Kaushik, et al. also examined different workflows to create QA datasets. They asked one set
of crowdworkers to write five questions after reading a passage, and another to write questions that elicit incorrect predictions from a pretrained QA model. Through this study, they derived insights about how each setup influenced crowdworker behavior, and then trained various QA models on these datasets.

Human-subjects research in NLP is not limited to studies aimed at dataset quality. Hayati, et al. paired two crowdworkers in a conversational setting and asked one to recommend a movie to the other. They analyzed the outputs to identify what communication strategies led to successful recommendations, and used these insights to train automated dialog systems.

Perez-Rosas, et al. asked crowdworkers to each write seven truths and seven plausible lies on topics of their own choosing, and collected demographic attributes (such as age and gender) for each crowdworker. They analyzed how attributes of deceptive behavior relate to gender and age, and then trained classifiers to predict deception, gender, and age. In these cases, the researchers interacted with crowdworkers to produce data about the crowdworkers that was then analyzed to answer research hypotheses, which created generalizable knowledge.

**Cases where the human-subjects designation is problematic**

Many ML crowdsourcing studies do not fit neatly in the paradigm of research common elsewhere. For example, crowdworkers are often recruited not as objects of study but to perform tasks that could have been—and sometimes are—performed by the researchers. In these cases, the
researchers interact with crowdworkers and produce data that is then used to produce generalizable knowledge. Moreover, some of the collected data is about the worker (e.g., to facilitate payment). In these cases, however, data analyzed to produce generalizable knowledge is not about the crowdworkers in any meaningful sense.

In the most common use of crowdsourcing in ML research (e.g., Hovy, et al.\textsuperscript{8}) workers are hired to label datasets used for model training. While such research might seemingly satisfy the interaction and use criteria from the Common Rule, it meets these through information not directly about the worker. Crowdworkers perform tasks that would typically be performed by the research team when dealing with smaller datasets. For example, Kovashka, et al. described computer vision papers where researchers provided their own labels.\textsuperscript{13} Addressing the same task, DeYoung, et al. recruited crowdworkers to provide annotation,\textsuperscript{3} while Zaidan, et al. did the annotations themselves.\textsuperscript{30} All these tasks involved interacting with crowdworkers and using the generated data.

On a strict reading of the claim that a human subject is a living individual “about whom” researchers obtain information that is used or analyzed to produce generalizable knowledge, crowdworkers in these cases would not be classified as human subjects. This reading is consistent with the practice of some IRBs. For example, Whittier College states:

Information-gathering interviews with questions that focus on things, products, or policies rather than people or their thoughts about themselves may not meet the definition of human-subjects
research. Example: interviewing students about campus cafeteria menus or managers about travel reimbursement policy.\textsuperscript{27}

In contrast, other IRBs adopt a far more expansive reading of the Common Rule. Loyola University says:

In making a determination about whether an activity constitutes research involving human subjects, ask yourself the following questions:

1) Will the data collected be publicly presented or published?

AND

2) Do my research methods involve a) direct and/or indirect interaction with participants via interviews, assessments, surveys, or observations, or b) access to identifiable private information about individuals, e.g., information that is not in the public domain?

If the answer to both these questions is “yes,” a project is considered research with human subjects and is subject to federal regulations.\textsuperscript{18}

Note that this interpretation does not distinguish whether the information is about an individual or just obtained via a direct and/or indirect interaction. This view appears to be shared by other IRBs as well.\textsuperscript{2}

How does information \textit{about} versus merely \textit{from} impact human-subjects determination? Traditionally, research ethics has not had to worry about who is a member of the research team and who is a research participant. This ambiguity arises in cases of self-experimentation, but such cases are rare and fit into the intervention + analysis category from the Common Rule. The scope of the effort
required to produce data that can be used in ML research has engendered new forms of interaction between researchers and the public. Without explicit guidance from federal authorities, individual IRBs have to grapple with this issue on their own.

Our contention is that in the problematic cases referred to in this section, crowdworkers are best understood as augmenting the labor capacity of researchers rather than participating as human subjects in that research. This argument has two parts.

The first part of the argument is based on symmetry. Within a division of labor, if a task can be performed by more than one person, the categorization of that task should depend on its substantive features, not the identity of the individual performing it. (The potential counterargument citing unionized and nonunionized workers or independent contractors and employees shows that individual identity and related features may influence workplace protections, even for the same type of work. Pre-existing agreements modifying agent entitlements, however, do not change the nature of the activity—be it work or research.)

Therefore, if the same task is performed by a researcher and then by crowdworkers, the categorization should be consistent across both instances. Consequently, symmetry implies that either both the crowdworker and the researcher are part of the research team or both are human subjects.

The second part of the argument offers additional factors encouraging the categorization of both as part of the research team. First, when conducting tasks pertinent
to the study, researchers do not self-experiment; they are not study subjects.

Second, this position reflects the understanding that these interactions generate useful information contributing to the development of universally applicable knowledge. This information, however, should be seen as originating from, not being about, them.

Third, researchers interact as a team to generate tools, materials, and metrics used in research. But this interaction and use creates the means of generating new knowledge; it does not constitute the data whose study or analysis produces new knowledge.

Finally, ignoring the distinction between data about a person versus from them and considering both researchers and crowdworkers as human subjects would excessively broaden the regulatory category. This would categorize every research team member, even in biomedical and social science, as human subjects since they regularly interact with their teams to generate information for general knowledge.

**Loopholes in research oversight**

Previous analysis has underscored an ethical quandary in ML research. Ethical oversight in studies involving human participants safeguards their interests, which can be at risk because of interactions, interventions, or subsequent data usage. The concept of an oversight loophole—where a researcher can evade oversight requirements without affecting the applied research procedures—constitutes an ethical concern. It infringes on the principle of equal treatment: If data is collected from individuals
for research to produce generalized knowledge, their interests should receive the same level of oversight and concern regardless of how labor is distributed during the process. Nevertheless, two aspects of ML research render it susceptible to oversight loopholes: (1) the way data collection and analysis workload is partitioned; and (2) how research questions often surface post-data collection.

SCENARIO 1

The Common Rule envisions several divisions of labor in research. In traditional biomedical or social science research, it is common for the same researchers to both collect and analyze data. This approach is affirmed by 45 CFR 46.102 (e)(1)(i), which states that a researcher who “[o]btains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens,” is engaging in human-subjects research. Here, the ethical review assesses whether (a) interactions respect participant autonomy and welfare, and (b) information obtained from these interactions is used in ways that respect individuals’ rights and welfare.

Data or biospecimens are often collected during medical care or other health services. Such interactions are governed by medical ethics and professional norms rather than requirements of research. Hence, research ethics review assesses if identifiable private information is contained in the data or specimens and whether its use respects individuals’ rights and welfare.

It is not clear whether the Common Rule accounts for cases where researchers collect data for research
goals but don’t analyze it themselves. This differs from secondary use of research data, where initial data collection already considers participants’ welfare and rights, ensuring adequate oversight. Subsequent oversight would thus evaluate additional use of that data.

In contrast, many ML researchers gather large datasets for research purposes, without defined hypotheses, often to support future research in broad fields. For example, Williams, et al. compiled a dataset for textual entailment recognition and released it (along with anonymized crowdworker identifiers) for future research. Similarly, Mihaylov, et al. and Talmor, et al. created and released QA datasets with anonymized identifiers for further research. These studies, involving only interaction with and using or analyzing data from crowdworkers, may not necessitate IRB review.

In a subsequent study, Geva, et al. analyzed information about crowdworkers using these anonymized sets. They assessed how ML models trained on data from one group of crowdworkers generalizes to data from another group, and trained models to predict the authoring crowdworkers for respective documents. Given that they studied only existing anonymous datasets and didn’t directly interact with the workers, it’s questionable whether their work would require IRB oversight. If the researchers who collected the initial data had also conducted this analysis, however, IRB review would have been compulsory to ensure proper protections of participants’ welfare.

While much of ML research poses minimal risk to participants, cases do exist where interventions or interactions are less benign. For example, Xu, et al. asked
crowdworkers to prompt unsafe responses from a chatbot, using this data to create safer response models. These individuals may not inherently be considered human subjects, as their input doesn’t pertain directly to them. In this study, however, the researchers also established an offensive language taxonomy for classifying human utterances, paving the way for its application in future research. Thus, inferences could potentially be drawn about the proclivities or proficiency of particular crowdworkers to use offensive language of particular types.

In each of these cases, datasets were collected that contain information from crowdworkers for the purposes of producing generalizable knowledge that can include information about the crowdworkers. A research oversight loophole is created as 45 CFR 46.102(e)(1)(i) considers individuals as human subjects only if their information is obtained and used in the same study. To be clear, releasing such a dataset with identifiable private information for research purposes would fall under clause (ii) from 45 CFR 46.102(e)(1) [discussed earlier in the section on the current regulatory framework]. Subsequent research on this dataset is also subject to this clause, as long as the identifiable information remains.

The method where researchers collect data from individuals to create generalizable knowledge, anonymize it, and pass it to another team for analysis could be viewed as a loophole. This process, unlike when the researchers themselves analyze the data, would not be subject to oversight aimed at respecting individual autonomy and welfare. Even though anonymization lessens harm from exposure of sensitive details, it doesn’t assure respect for
individual autonomy and well-being in the data collection, due to absence of oversight.

One way to address loopholes of this type would be to amend 45 CFR 46.102 (e)(1)(i) to explicitly include the release of data alongside its use, study, or analysis.

SCENARIO 2
Revision of 45 CFR 46.102 (e)(1)(i) to include data release might not prevent loopholes. For example, a research team collecting data directly from crowdworkers and about them—an approach fitting standard research—might divide the process into two protocols to avoid IRB approval requirements. In the first protocol, they collect data but analyze only that from the crowdworkers and not about them. They then anonymize all collected data and in a second protocol analyze data about crowdworkers. This avoids the need for IRB approval as it doesn’t involve personal interaction or identifiable private information use.

In this scenario, a single study that would require IRB approval could avoid research ethics oversight by being decomposed into separate studies. As a result, the determination of whether an ML project constitutes research with human participants might need to be made at a higher level than the individual study protocol.

In the context of drug development, for example, a trial portfolio has been defined as a “series of trials that are interrelated by a common set of objectives.” It might be beneficial to apply this portfolio-level approach in ML research—that is, considering the data generated and questions investigated across interlinked studies’ relevance to crowdworkers. Successful portfolio-level
reviews need researchers to predetermine the kind, scope, and nature of data they are collecting and possible inquiries across various studies. As new research questions often arise after data collection because of the dynamic nature of ML research, researchers may need to consult with IRBs to clarify when a proposed portfolio of studies should be classified as human research.

DISCUSSION
There is considerable confusion about when ML’s crowdworkers constitute human subjects for ethical and regulatory purposes. While some sources suggest treating all crowdworkers as human subjects,23 our analysis makes a more nuanced proposal, identifying: (1) clear-cut cases of human-subjects research, which require IRB consultation, even if only to confirm that they belong to an exempt category; (2) crowdsourcing studies that do not constitute human-subjects research because the analyses do not involve data about the workers; (3) difficult cases, where the distinctive features of ML’s crowdworking studies combine with ambiguities in the Common Rule to create uncertainty about how to apply existing requirements; and (4) loopholes, whereby researchers might elude the human-subjects designation without making substantive changes to the research performed.

The spirit of research oversight is to safeguard the rights and interests of individuals involved in research. Individuals who are not research participants can still be exposed to risks to their well-being and threats to their autonomy. This is particularly true of employment interactions, as employers often have access to sensitive, private,
identifiable information (such as Social Security Number and background-check reports) about their employees.

The solution is not necessarily to redefine all crowdworkers as human subjects, but rather to clarify the parameters for their classification as such, ensuring due oversight when applicable. In other instances, their rights should be upheld via ethical and regulatory frameworks guiding labor practices and workplace safety.

Our recommendations

- ML researchers must work proactively with IRBs to determine which, if any, information they will generate is about versus merely from crowdworkers. They must discern whether their intended portfolio of studies involving this data constitutes human-subjects research. They should also recognize that as the questions they investigate change, the status of the research they are conducting may change. Consequently, researchers must consult IRBs to understand when a new submission or a protocol modification is necessary for the ongoing research.

- IRBs should not reflexively classify all ML research involving crowdworkers as human-subjects research. Rather, IRBs should establish clear procedures for evaluating portfolios of research to address the possibility of loopholes in research oversight. They should communicate with ML researchers about the conditions under which the classifications might change and the conditions under which a revised protocol would be required.

- OHRP should offer precise guidance about what it
means for information or analysis to be “about” a set of individuals. We also recommend that OHRP revise the Common Rule so that 45 CFR 46.102(e)(1) condition (i) reads: “Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, analyzes, or releases the information or biospecimens.” This modification would require that an original investigator who collects data through interaction with humans and plans to release a dataset (even if anonymized) that could be used to ask questions about those individuals must secure IRB approval for the research in which that data is gathered. Subsequent studies using the anonymized data would not be counted as human-subjects research unless they aim to re-identify individuals. This change resolves one loophole identified here. OHRP also has a role to play in offering guidance to ML researchers. This could be achieved by issuing an agency Dear Colleague letter or an FAQ document.

Acknowledgments
The authors thank Sina Fazelpour, Holly Fernandez Lynch, and I. Glenn Cohen for their constructive feedback. They also thank Mozilla, the Carnegie Mellon University Block Center, the Carnegie Mellon University PwC Center, University of Pittsburgh Medical Center, Abridge, Meta Research, and Amazon Research for the grants and fellowships that made this work possible.

References
1. Adelani, D. F., Abbott, J., Neubig, G., D’souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H.,
2. Birmingham-Southern College. Do I need IRB approval? https://www.bsc.edu/academics/irb/documents/BSC%20IRB%20Decision%20Tree.pdf.

3. DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., Wallace, B. C. 2020. ERASER: a benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4,443–4,458; https://aclanthology.org/2020.acl-main.408/.

4. Dodge, J., Gururangan, S., Card, D., Schwartz, R., Smith, N. A. 2019. Show your work: improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2,185–2,194; https://aclanthology.org/D19-1224/.

5. Fort, K., Adda, G., Cohen, K. B. 2011. Amazon Mechanical Turk: gold mine or coal mine? *Computational Linguistics* 37(2), 413–420; https://aclanthology.org/J11-2010.pdf.

6. Geva, M., Goldberg, Y., Berant, J. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1,161–1,166; https://aclanthology.org/D19-1107.pdf.
7. Hayati, S. A., Kang, D., Zhu, Q., Shi, W., Yu, Z. 2020. INSPIRED: toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8,142–8,152; https://aclanthology.org/2020.emnlp-main.654/.

8. Hovy, D., Plank, B., Søgaard, A. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 377–382; https://aclanthology.org/P14-2062.pdf.

9. Hovy, D., Spruit, S. L. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598; https://aclanthology.org/P16-2096/.

10. Ipeirotis, P. 2009. Mechanical Turk, human subjects, and IRBs; https://www.behind-the-enemy-lines.com/2009/01/mechanical-turk-human-subjects-and-irbs.html.

11. Kaushik, D., Kiela, D., Lipton, Z. C., Yih, W.-T. 2021. On the efficacy of adversarial data collection for question answering results from a large-scale randomized study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 6,618–6,633; https://aclanthology.org/2021.acl-long.517.pdf.

12. Khashabi, D., Khot, T., Sabharwal, A. 2020. More bang for your buck: natural perturbation for robust question
answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 163–170; https://aclanthology.org/2020.emnlp-main.12.pdf.

13. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K. 2016. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision* 10(3), 177–243; https://www.nowpublishers.com/article/Details/CGV-071.

14. Leidner, J. L., Plachouras, V. 2017. Ethical by design: ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 30–40; https://aclanthology.org/W17-1604.pdf.

15. London, A. J.. 2021. *For the Common Good: Philosophical Foundations of Research Ethics*. Oxford University Press.

16. London, A. J., Kimmelman, J. 2019. Clinical trial portfolios: a critical oversight in human research ethics, drug regulation, and policy. *Hastings Center Report* 49(4), 31–41; https://pubmed.ncbi.nlm.nih.gov/31429954/.

17. London, A. J., Taljaard, M., Weijer, C. 2020. Loopholes in the research ethics system? Informed consent waivers in cluster randomized trials with individual-level intervention. *Ethics & Human Research* 42(6), 21–28; https://onlinelibrary.wiley.com/doi/abs/10.1002/ehar.500071.

18. Loyola University. Do I need IRB review? https://web.archive.org/web/20220214194123/https://www.whittier.edu/academics/researchethics/irb/need.

19. Mihaylov, T., Clark, P., Khot, T., Sabharwal, A. 2018. Can a suit of armor conduct electricity? A new dataset
for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2,381–2,391; http://aclanthology. lst.uni-saarland.de/D18-1260.pdf.

20. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F., et al. 2020. Participatory research for low-resourced machine translation: a case study in African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2,144–2,160; https://aclanthology.org/2020.findings-emnlp.195.pdf.

21. Ning, Q., Wu, H., Dasigi, P., Dua, D., Gardner, M., Logan IV, R. L., Marasovic, A., Nie, Z. 2020. Easy, reproducible, and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Empirical Methods in Natural Language Processing (EMNLP), Systems Demonstrations*, 127–134; https://aclanthology.org/2020.emnlp-demos.17.pdf.

22. Pérez-Rosas, V., Mihalcea, R. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1,120–1,125; https://aclanthology.org/D15-1133.pdf.

23. Shmueli, B., Fell, J., Ray, S., Ku, L.-K. 2021. Beyond fair pay: ethical Implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3,758–3,769; https://aclanthology.org/2021.naacl-main.295.pdf.
24. Talmor, A., Herzig, J., Lourie, N., Berant, J. 2019. CommonsenseQA: a question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4,149–4,158; https://aclanthology.org/N19-1421.pdf.

25. Vaughan, J. W. 2017. Making better use of the crowd: how crowdsourcing can advance machine learning research. Journal of Machine Learning Research 18 (1), 7,026–7,071; https://dl.acm.org/doi/10.5555/3122009.3242050.

26. Whiting, M. E., Hugh, G., Bernstein, M. S. 2019. Fair work: crowd work minimum wage with one line of code. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7, 197–206; https://ojs.aaai.org/index.php/HCOMP/article/view/5283.

27. Whittier College. Do I need IRB Review? https://web.archive.org/web/20220214194123/https://www.whittier.edu/academics/researchethics/irb/need.

28. Williams, A., Nangia, N., Bowman, S. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (Long Papers), 1,112–1,122; https://aclanthology.org/N18-1101.pdf.

29. Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., Dinan, E. 2020. Recipes for safety in open-domain chatbots.
arXiv preprint arXiv:2010.07079; https://arxiv.org/abs/2010.07079.

30. Zaidan, O., Eisner, J., Piatko, C. 2007. Using “Annotator Rationales” to improve machine learning for text categorization. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, 260–267; https://aclanthology.org/N07-1033.

31. Zhang, D., Zhang, M., Zhang, H., Yang, L., Lin, H. 2021. MultiMET: A multimodal dataset for metaphor understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 1: Long Papers, 3,214–3,225; https://aclanthology.org/2021.acl-long.249.pdf.

Divyansh Kaushik is the associate director for emerging technologies and national security at the Federation of American Scientists, where his work focuses on AI policy and U.S.-China strategic competition. He holds a Ph.D. from Carnegie Mellon University, where he focused on designing reliable artificial intelligence systems that align with human values.

Zachary C. Lipton is an assistant professor of machine learning at Carnegie Mellon University (CMU) and the chief technology officer and chief scientist at the healthcare startup Abridge. At CMU, he directs the Approximately Correct Machine Intelligence (ACMI) lab, whose research focuses include the theoretical and engineering foundations of robust and adaptive machine learning algorithms,
applications to both prediction and decision-making problems in clinical medicine, natural language processing, and the impact of machine learning systems on society. A key theme in his current work is to take advantage of causal structure underlying the observed data while producing algorithms that are compatible with the modern deep learning power tools that dominate practical applications. He is the founder of the Approximately Correct blog and a co-author of Dive into Deep Learning, an interactive open-source book drafted entirely through Jupyter notebooks that has reached millions of readers. He can be found on Twitter (@zacharylipton), GitHub (@zackchase), or his lab’s website (acmilab.org).

**Alex John London** is the K&L Gates Professor of Ethics and Computational Technologies, co-lead of the K&L Gates Initiative in Ethics and Computational Technologies at Carnegie Mellon University, director of the Center for Ethics and Policy at Carnegie Mellon University, and chief ethicist at the Block Center for Technology and Society at Carnegie Mellon University. An elected Fellow of the Hastings Center, Professor London’s work focuses on ethical and policy issues surrounding the development and deployment of novel technologies in medicine, biotechnology, and artificial intelligence. His book, For the Common Good: Philosophical Foundations of Research Ethics, is available in hard copy from Oxford University Press and as an open access title.