Global Journal of Engineering and Technology [GJET].
ISSN: 2583-3359 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journal-gjet-home/

PROGRESSING TOWARDS PARTICIPATORY AI: A SAFETY EVALUATION FRAMEWORK

BY

Ayse Kok Arslan

Oxford Alumni-Silicon Valley

Article History

Received: 08/10/2023
Accepted: 18/10/2023
Published: 20/10/2023

Vol –2 Issue – 10

PP: - 20-24

Abstract

This study explores main problem areas that would help make progress on participatory AI models; namely, robustness, monitoring, alignment, and systemic safety. For each of the four problems, it discusses possible research directions and provides an overview of how to guard against extreme risks while developing and deploying a model. It also identifies new problems, such as emergent capabilities from massive pre-trained models grounded in recent progress in participatory AI models. It concludes that agency is an important property to evaluate given the central role of agency in various theories of AI risk.

Keywords: *AI, fairness, machine learning, algorithms, computing, tech, software design*

Introduction

As AI progress has advanced, general-purpose AI systems evolved to display both new and harmful capabilities that their developers did not intend. A central goal of AI governance should be to limit the creation, deployment, and proliferation of systems that pose extreme risks.

This study focuses on paths towards creating safe AI systems by exploring main problem areas that would help make progress on participatory AI models, robustness, monitoring, alignment, and systemic safety. Within this context, AI Safety research aims at making the adoption of AI more beneficial by focusing on long-term and long-tail risks.

An inquiry into the emergent capabilities of modern AI systems also broadens the scope by identifying systemic safety risks surrounding the deployment context of AI.

Review of Existing Work

Broadly, public participation refers to approaches or activities that engage or involve members of the public, incorporating perspectives and experiences into a project or intervention. Participation has also connotations with a 'moral good' [46], or 'flourishing social ties'[1].

While in the domain of policy, the 'public' may refer to 'citizens', 'labeling data people', or 'lay persons' [4], it may refer to current or future 'end users' [6] in the context of technology.

More recent literature around participation in AI adopts a broader definition that includes all people affected by the use of an AI system, particularly individuals and groups for whom AI risks exacerbating inequity, injustice as well as marginalization [7].

Yet, given the conceptual confusion about 'participation' in AI in the existing literature, there is a lack of understanding of what kinds of approaches should be adopted [8, 2]. This could eventually slow down the adoption of these models.

The form of public participation can vary, as reflected in the various typologies produced by scholars and practitioners [2, 5]. Two existing typologies are instructive for classifying the different modes of participation in AI:

- Sloane et al.'s typology of participation as work, as consultation, and as justice [77],

- Birhane et al.'s exploration of the three instrumental categories of participation for algorithmic performance improvement; for process improvement and for collective exploration [8].

Based on these initial frameworks, there is an emerging literature on participatory approaches to AI development, which identify a few kinds of 'participatory' activities that involve assembling a mixed group of stakeholders to consult or assess an AI system.

The first of these is Sherry Arnstein's Ladder of Citizen Participation [2]- a widely referenced framework for forms of participation- which is originally intended to outline different degrees of participatory approaches in public planning.

Accordingly, there are eight levels ranging from forms of non-participation (‘manipulation’), one-way dialogic methods (such as public request for comment [5]), involvement by consultation and partnership in the middle layers, and finally ‘citizen control’ at the top level (see Figure 1). Arnstein criticizes these approaches and describes them as being tokenistic and inadequate in shifting the axis of power.

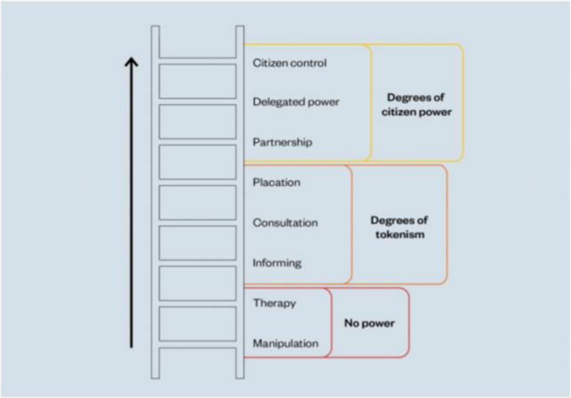

Fig.1 Arnstein’s Framework for Citizen Participation

Patel et al. [6] drew on Arnstein’s ladder and a more recent ‘spectrum of participation’ [5] to describe practical mechanisms of participation and consequently the design of data-driven systems, including AI (see Figure 2).

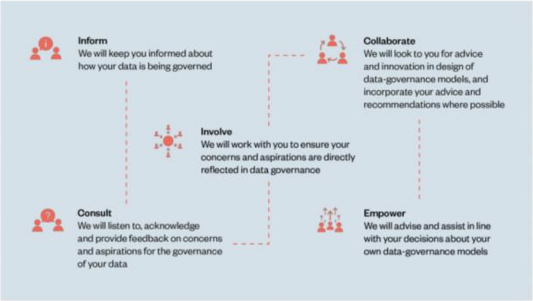
Inform
We will keep you informed about how your data is being governed
Collaborate
We will look to you for advice and innovation in design of data-governance models, and incorporate your advice and recommendations where possible
Involve
We will work with you to ensure your concerns and aspirations are directly reflected in data governance
Consult
We will listen to, acknowledge and provide feedback on concerns and aspirations for the governance of your data
Empower
We will advise and assist in line with your decisions about your own data-governance models

Fig.2 Spectrum of Participation (Patel et. al, 2018)

This analysis provides five levels of participation and examples of what practical mechanisms may exist for each, drawn from real-world case studies. These five levels include:

1. Informing people about how data about them is used, such as through the publication of model cards;
2. Consulting people to understand their needs and concerns in relation to data use, such as through user experience research or consumer surveys;
3. Involving people in the governance of data, such as through public deliberation or lived experience panels;
4. Collaborating with people in the design of data governance structures;
5. Empowering people to make decisions about datasets and technologies built with them, such as through citizen-led governance boards.

These taxonomies help AI stakeholders makes sense of the public participation approaches AI companies may be using and contribute theoretical foundational frameworks for exploring participation in AI design.

Activities such as crowdsourcing [3,8] data labelling[8], creating ‘red teams’ to test or evaluate a model [4], or engaging members of the public to elicit preferences for algorithmic design decisions [12, 7] are usually seen as being ‘participatory’. Nevertheless, such forms of participation very often prioritize a higher total number of participants over the length or depth of participant involvement [5].

Furthermore, it is also possible that harnessing AI’s potential for global benefit and managing its risks could require participation of governing entities at the international level. According to a recent paper by DeepMind (2023), in order to make international institutions participate in AI research, institutional functions can be grouped into the following categories (4, 7, 9):

- *Conduct or support AI safety research:* This includes the measures to reduce the risks of AI misuse by means of training to manage risky behaviors, and examining safe deployment protocols appropriate to different system [3 20].
- *Distribute and enable access to cutting-edge AI:* This refers to facilitating access to cutting-edge systems and increasing absorptive capacity through education, infrastructure, and support of the local commercial ecosystem.
- *Set safety norms and standards:* This includes guidelines and standards around how AI can be developed, deployed, and regulated to maximize benefits and minimize risks.
- *Monitor compliance:* This entails audits of issue certifications and evaluations to ensure adherence to international standards and agreements.
- *Control AI inputs:* This includes how to monitor models, compute, data, and other ingredients of potentially dangerous technologies.

Other scholarship argues that participatory approaches in AI could be instrumentalized to advance ambitious societal-level goals such as fairness, inclusion [9,11], justice [6, 17], and accountability [12] which could be characterized as Sloane et al.’s ‘participation as justice’ [77].

There are also dangers, as noted by Lloyd et al., that a focus on engaging technology ‘users’ in participatory projects could narrow focus away from broader segments of society that might be affected by AI, with a risk of exacerbating existing harms to these groups [17]. Tech companies could also use ethics initiatives as a form of social capital that justifies de-regulation of their industry in favor of self-regulation.

To better harness advanced AI for global benefit, international efforts to help underserved societies access and use advanced AI systems will be important.

A summary review of existing studies with regard to the public participation in AI provides the following results:

1. Within commercial AI settings, public participation is viewed as serving societally ‘good’ ends, but may also have a strong business purpose.
2. Public participation in AI industry lacks clear and shared understanding of practices. Participants did not identify many participatory methods they use but rather tended to list methods they had heard of.
3. Public participation in AI labs faces various obstacles: resource-intensity, atomization, exploitation risk, and mis-aligned incentives.
4. Public participation in AI is complicated by-products or research that lack a clear context.

An early work that helps identify safety problems is Russell et al., 2015 [154], who identify many potential avenues for safety, spanning robustness, machine ethics, research on AI’s economic impact, and more.

Amodei and Olah et al., 2016 [5] helped further explore several safety research directions.

These scholars highlight the importance of various other research problems including adversarial training and uncertainty estimation. Certain types of research, such as the development of model evaluations can proceed effectively with API access to the models, while others such as mechanistic interpretability might require access to the model weights and architectures [17].

In a similar vein, Vaswani et al. (2017) asserted in many models, the control flows are specified by hidden weights through means of gradient optimizers rather than being programmed via means of modularity or encapsulation. Therefore, there is a need to train new position embeddings or global attention weights as well.

While many effective attacks assume full access to a neural network, sometimes assuming limited access is more realistic. If a black box system is not publicly released and can only be queried, it may be possible to practically defend the system against zero-query attacks [9] or limited-query attacks [5].

Another more white-box approach would be to predict a model’s capabilities given only its weights, which might reveal latent capabilities that are not obviously expressible from standard prompts.

Training models will also need to adapt to an evolving world given the high level of uncertainty and be able to be improved based on novel experiences [1, 6, 8]. To guard against extreme risks, AI developers should use model evaluation to uncover:

1. To what extent a model is capable of causing extreme harm (which relies on evaluating for certain dangerous capabilities).
2. To what extent a model has the propensity to cause extreme harm (which relies on alignment evaluations).

AI models could also help predict future phases of cyberattacks, and such automated warnings could be judged by their lead time, precision, recall, and the quality of their contextualized explanation.

Future systems will operate in environments that are broader, larger-scale, and more highly connected with more feedback loops, paving the way to more extreme events [13] than those seen today.

Eventually, advisory systems could identify stakeholders, propose metrics, brainstorm options, suggest alternatives, and note trade-offs to further improve decision quality [58]. In summary, AI systems can help prevent incidents arising due to jumping to conclusions or reduce inadvertent escalations.

Recommended Framework

As seen in Figure 3., AI developers and regulators must be able to identify AI capabilities to limit the risks they pose. The AI community already relies heavily on model evaluation – i.e. empirical assessment of a model’s properties – for identifying and responding to a wide range of risks.

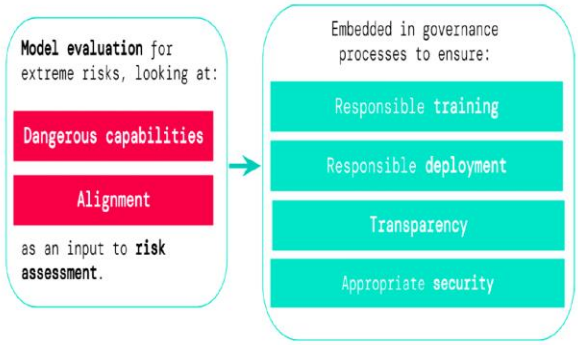
Model evaluation for extreme risks, looking at:
Dangerous capabilities
Alignment
as an input to risk assessment.
Embedded in governance processes to ensure:
Responsible training
Responsible deployment
Transparency
Appropriate security

Figure 3. A Change for Model Evaluations

According to the framework displayed in Figure 3, evaluations should ensure the following:

1. *Responsible training:* Responsible decisions are made about whether and how to train a new model that shows early signs of risk.
2. *Responsible deployment:* Responsible decisions are made about whether, when, and how to deploy potentially risky models.
3. *Transparency:* Useful and actionable information is reported to stakeholders, to help them mitigate potential risks.
4. *Appropriate security:* Strong information security controls and systems are applied to models that might pose extreme risks.

Model evaluation is only one among several tools available for AI risk assessment – more theoretical approaches are also available. These evaluations can be organized into two categories:

(a) whether a model has certain dangerous capabilities, and
(b) Whether it has the propensity to harmfully apply its capabilities (alignment).

Alignment evaluations should look for behaviors such as whether the model:

- Pursues long-term, real-world goals, different from those supplied by the developer or user (Chan et al., 2023; Ngo et al., 2022);
- Engages in “power-seeking” behaviors (Krakovna and Kramar, 2023; Turner et al., 2021);
- Resists being shut down (Hadfield-Menell et al., 2016; Orseau and Armstrong, 2016);
- Can be induced into collusion with other AI systems against human interests (Ngo et al., 2022).

Figure 4 provides an overview of how to guard against extreme risks while developing and deploying a model, with evaluation embedded throughout. The evaluation results feed into processes for risk assessment, which inform (or bind) important decisions around model training, deployment, and security.

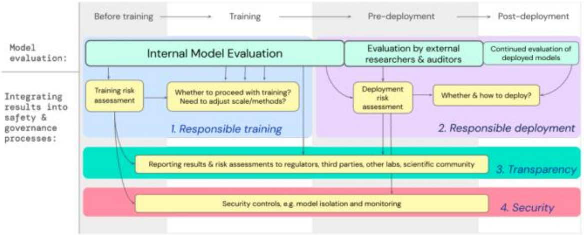
Before training
Training
Pre-deployment
Post-deployment
Model evaluation:
Integrating results into safety & governance processes:
Internal Model Evaluation
Evaluation by external researchers & auditors
Continued evaluation of deployed models
Training risk assessment
Whether to proceed with training? Need to adjust scale/methods?
Deployment risk assessment
Whether & how to deploy?
1. Responsible training
2. Responsible deployment
Reporting results & risk assessments to regulators, third parties, other labs, scientific community
3. Transparency
Security controls, e.g. model isolation and monitoring
4. Security

Figure 4. Workflow for model deployment

Three sources of model evaluations feed into this process:

1. Internal model evaluation: Internal researchers have usually high context on the model’s design and deeper model access than can be achieved via an API. Developers could establish an internal safety evaluation function that is independent of the teams primarily responsible for building the models, reporting directly to organizational leaders (see Raji et al., 2020).
2. External research access: The developer grants model access to external researchers, likely via an API (Bluemke et al., 2023; Shevlane, 2022a,b). Their research could be exploratory or targeted at evaluating specific properties, including “red teaming” the model’s alignment.
3. External model audit, i.e. model evaluation by an independent, external auditor for the purpose of providing a judgement — or input to a judgement — about the safety of deploying a model (or training a new one) (ARC Evals, 2023; M.kander et al., 2023; Raji et al., 2022b).

Ideally there would exist a rich ecosystem of model auditors providing broad coverage across different risk areas. (This ecosystem is currently underdeveloped.)

Responsible Training

Before a frontier training run, developers have the opportunity to study weaker models that might provide early warning signs. These models come from two sources:

(1) previous training runs, and

(2) Experimental models leading up to the new training run.

Developers should evaluate these models and try to forecast the results from the planned training run (see OpenAI, 2023b). The developer has a range of possible responses to address the concerning evaluation results:

1. *Study the issue:* This in essential to understand why the misalignment or dangerous capability emerged.
2. *Adjust the training methods to circumvent the issue:* This could mean adjusting (for example) the architecture, the data, the training tasks, or further developing the alignment techniques used.
3. *Careful scaling*: If the developer is not confident it can train a safe model at the scale it initially had planned, they could instead train a smaller or otherwise weaker model.

Responsible Deployment

Deployment means making the model available for use, e.g. it is built into a product or hosted on an API for software developers to build with.

Model evaluation for extreme risks could inform a deployment risk assessment that reviews (a) whether or not the model is safe to deploy, and (b) the appropriate guardrails for ensuring the deployment is safe.

In response to concerning evaluation results, one possibility is to recommend against deployment.

A second possibility is to recommend adjustments to the deployment plan that would address potential risks.

Responsible Audit

Evaluation will often need to continue after deployment. There are two reasons for this:

1. Unanticipated behaviors: Before deployment, it is impossible to fully anticipate and understand how the model will interact in a complex deployment environment. Therefore, in the early stages of deployment, developers must:

(a) Monitor emerging model behaviors and risks: These efforts include direct monitoring of inputs and outputs to the model, and systems for incident reporting (see Brundage et al., 2022; Raji et al., 2022b).

(b) Design and run new model evaluations inspired by these observations.

2. Updates to the model: The developer might update the model after deployment, e.g. by fine-tuning on data collected during deployment or by expanding the model’s access to externaltools.

Recommendations

Within the light of this information, researchers must evaluate a model across a broad range of settings by taking into account the following factors:

1. Breadth: Evaluating behavior across as wide a range of settings as possible. One promising avenue is automating the process of writing evaluations using AI systems (Perez et al., 2022b) (see also Pan et al., 2023).

2. Targeting: Some settings are much more likely to reveal alignment failures than others, such as using gradient-based adversarial testing and related approaches (Jones et al., 2023) which may be more beneficial for everyone.
3. Understanding generalization: Since researchers will be unable to foresee or simulate all possible scenarios, there is a need to develop a better scientific understanding of how and why model behaviors generalize (or fail to generalize) between settings.

Last, but not least, agency – in particular, the goal-directedness of an AI system – is an important property to evaluate (Kenton et al., 2022), given the central role of agency in various theories of AI risk (Chan et al., 2023). Partly, agency is a question of the model's capabilities – is it capable of effectively pursuing goals. Evaluating alignment also requires looking at agency.

Conclusion

Public participation is recognized as a valuable mechanism to involve public perspectives which is viewed as a way to mitigate risks in AI systems and produce more 'societally beneficial' technologies.

This study explored current conditions and emergent challenges for public participation in commercial AI to lay foundations for further work and debate. It also provided an overview of how to guard against extreme risks while developing and deploying a model with a focus on agency as an important property in various theories of AI risk.

Successful AI requires a clear use case for members of the public to understand, raising an innate challenge for the use of these methods for general-purpose technologies.

References

1. Akhtar: Google defends its search engine against charges it favors Clinton,‖ USA Today (10 June) https://www.usatoday.com/story/tech/news/2016/06/10/google-says-search-isntbiased-toward-hillaryclinton/85725014/, accessed 14 July 2020 (2016)
2. Arentz, W and B. Olstad: Classifying offensive sites based on image content,‖ Computer Vision and Image Understanding, volume 94, numbers 1–3, pp 295– 310. doi https://doi.org/10.1016/j.cviu.2003.10.007, accessed 14 July 2020. (2016).
3. Gulli, A: A deeper look at Autosuggest,‖ Microsoft Bing Blogs (25 March), at https://blogs.bing.com/search/2013/03/25/a-deeper-look-at-autosuggest/, accessed 14 July 2020. (2013)
4. McGuffie and A. Newhouse: The radicalization risks of GPT-3 and advanced neural language models,‖ arXiv:2009.06807v1 (15 September), at https://arxiv.org/abs/2009.06807, accessed 9 April 2021. (2020)
5. Miller and I. Record, M: Responsible epistemic technologies: A social-epistemological analysis of autocompleted Web search,‖ New Media & Society, volume 19, number 12, pp. 1,945–1,963. doi: https://doi.org/10.1177/1461444816644805, accessed 14 July 2020. (2017)
6. Olteanu, C. Castillo, J. Boy, and K. Varshey: The effect of extremist violence on hateful speech online,‖ Proceedings of the Twelfth International AAAI Conference on Web and Social Media, at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17908/17013, accessed 14 July 2020 (2018)
7. Olteanu, K. Talamadupula, and K. Varshney: The limits of abstract evaluation metrics: The case of hate speech detection,‖ WebSci '17: Proceedings of the 2017 ACM on Web Science Conference, pp. 405–406.doi: https://doi.org/10.1145/3091478.3098871, accessed 30 January 2022. (2017)
8. Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil: Learning semantic representations using convolutional neural networks for Web search,‖ WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web, pp. 373– 374. doi: https://doi.org/10.1145/2567948.2577348, accessed 14 July 2017.
9. Olteanu, C. Castillo, F. Diaz, and E. Kcman: Social data: Biases, methodological pitfalls, and ethical boundaries,‖ Frontiers in Big Data (11 July).doi: https://doi.org/10.3389/fdata.2019.00013, accessed 14 July 2020. (2019)
10. H. Yenala, M. Chinnakotla, and J. Goyal: Convolutional bi-directional LSTM for detecting inappropriate query suggestions in Web search,‖ In J. Kim, K. Shim, L. Cao, J.G. Lee, X. Lin, and Y.S. Moon (editors). Advances in knowledge discovery and data mining. Lecture Notes in Computer Science, volume 10234. Cham, Switzerland: Springer, pp. 3–16.doi: https://doi.org/10.1007/978-3-319-57454- 7_1, accessed 14 July 2020. (2017)