# Dataset Reflection Questions (DRQs)

Ethics Review Program and Responsible AI Program
Microsoft Research, Technology & Research Division
Last updated: June 2025

## Instructions

These questions are meant to stimulate thinking around the potential impacts of using a given dataset for research, and to steer decisions you make about either how to conduct the research or how you communicate the strengths/limitations of your findings to your intended audience. Be sure to still connect with your Institutional Review Board for official reviews and consultation, as appropriate.

**How should you approach this exercise?** Every research project will have unique considerations. Use both the questions below and your subject matter expertise to guide you. It's always helpful to get additional perspectives. Not all questions will apply to all datasets or data uses.

You may not know all the answers to the questions below. Think about how what you don't know about the dataset – e.g., its exact contents or how it was created – might affect how you use the dataset going forward. If you're working with multiple datasets, think about both the individual characteristics of each dataset and how their use together affects your answers to these questions.

**Think about impacts as you work through each question.** Consider impacts across three main stakeholder categories:

- Impacts to *specific individuals that contributed to the data*.
- Impacts to *groups that are represented by the data or are absent from the data*.
- Impacts to *wider society*.

Consider impacts for both **immediate** and **future** uses. Prioritize immediate impacts now, as you may not know all the use scenarios for your work yet. Where possible, capture your thoughts about potential impacts for future uses, including:

- What happens when you share your research results?
- What happens if/when your work is widely adopted?

# Reflection Questions

## Representativeness

1. What do you know about how inclusive the dataset is across various populations, geographies, or other characteristics related to your research question?

   a) What populations, languages, or geographies might be excluded from the data altogether?
   b) Where is there likely to be over or under representation?
   c) How could the representation of the population in the dataset affect how well any models built from this dataset work for populations not represented in the dataset?

2. Might this dataset have potential for over-generalizing, stereotyping, or perpetuating falsehoods about subpopulations/minority characteristics/behavior/views/perspectives? (e.g., the dataset is text from public forums where stereotyping language is routinely used.)

   a) What specific fairness-related issues are most likely?

3. Is there anything about the composition of the dataset, or the way it was collected/processed/cleaned/labeled, that might impact the suitability of the dataset for your use? For example, the original domain of the data doesn't match the domain of your use.

4. Given the makeup of this data, what tasks or scenarios would the data be particularly well or ill-suited for?

5. Who paid for the data? Who funded the creation of this dataset? How might that affect the makeup or quality of the dataset? What conflicts of interest could there be with the makeup of the data given who funded it?

6. Is there a different dataset, or a combination of datasets, that might be more representative/inclusive?

## Risks versus Benefits

7. What are the risks to each stakeholder group? (Weigh these against benefits)

8. What are the benefits to each stakeholder group? (Weigh these against risks)

9. How are the risks to each stakeholder group distributed? Is the distribution uneven across any vulnerable populations (e.g., children, elderly, prisoners) or historically marginalized groups?

10. Consider how the study design might be altered to minimize disparate impacts and/or how you might disclose the limitations of this study / areas for future research at the time of final writeup.

## Considerations for Stakeholder Risks

11. What might research results imply or suggest about the people/communities/groups represented in the data? Could this harm anyone?

12. What possibility is there that this dataset contains potentially offensive or illegal content? (e.g. violent imagery, pornography, racist language, hate speech, etc.)

a) How might that affect any research subject interacting with the data (e.g. annotation)?
b) Looking back at your assessment of identifiability (along a spectrum), what more can you do to minimize the risk of disclosing sensitive/incriminating information about someone?

13. For datasets involving search terms entered by humans:

a) Be deliberate about the list of key words that you are looking for in search text, so as not to capture more than you need.
b) Why have you chosen these specific keywords for your study? How representative are the terms you have chosen for the concepts you are testing in your research?
c) Would any search terms reasonably be considered sensitive - e.g., pertaining to health status, sexual behavior or orientation, immigration/citizenship status, employment status, criminal behavior, etc.? Consider not only what conclusions around sensitive topics might say about individuals (who may be anonymized), but also whether your results might be interpreted to describe a class of people, like a neighborhood, income bracket, race/ethnicity, etc. Was this your intention?
d) Will any metadata accompany the search terms, such as IP address, zip code, time-to-click, hover time, etc.?

## Identifiability

14. How identifiable is it? When responding to these questions, consider how identifiability might be more of a spectrum, rather than a binary yes/no.

15. Are there any direct identifiers in the data (e.g., name, email address, audio or video recordings) or information that might be combined to identify someone, like age, gender, sexual orientation, place of employment, telemetry, etc.?

16. Is the data linkable to identifiers using IDs, codes or pseudonyms? If so, who has access to the key? Are there any agreements prohibiting your team from accessing the key?

17. Are there any free text fields in the data in which an individual might knowingly or unknowingly identify themselves or third parties? (e.g., chats)

18. What is the estimated total number of subjects represented in the data? How does that affect identifiability?

19. How likely is it that individuals could be identified by unique characteristics (e.g. rare disease, tattoos, census track) or by combining characteristics?

20. Do you or other researchers need to identify people to address your research question? Is there a scientific benefit or any advantage to doing so? When in doubt, collect no/fewer identifiers.

21. Could any alterations/analysis performed by you or other researchers, including merging datasets, make the data more identifiable than it was or reveal more information about subjects than was otherwise present in the data? If so, how does this affect risks to participants or groups?

## Consent and Expectations

22. How accessible is the data? Can anyone access the data for free or is access to this data restricted in any way? For example, must you be a member of a professional or social organization? Must you subscribe? Must you complete training? This speaks to how "public" the dataset is, and may be an indicator of expectation of privacy among people whom the data is about.

23. You may have already considered how people were informed about the use of their data or how they gave consent. Here is a more comprehensive set of questions about consent to provoke deeper thought, not just about privacy law, but also about showing respect to individuals who contributed to your dataset.

a) Did the people represented in the data know their information was collected?
b) Did they explicitly agree to collection (e.g., click or sign a consent form)?
c) Did they passively agree to collection (e.g., a terms of service allows the use of their data)?

d) Did they agree to making their data available *in this way*, either explicitly or passively?

e) Is it reasonable to think they know their information is currently available for your intended purpose?

24. Is the intended use compatible with whatever the participants agreed to, if anything?

25. Would people who contributed to the dataset support your reasoning for reusing their data for this research?

26. How well aligned is your use with the original reasons for collecting the data (e.g., hospital records created for someone admitted to surgery)?

## Transparency

27. Knowing what you know about the makeup of your data, what should others know about your work when reading your paper or using the technology you developed?

28. What are the best use cases, given the dataset you used, and what uses should be either avoided or studied further?

29. Who do your findings/tool best represent/serve?

30. Who was not represented in the data that your research is based on? And how does this impact the generalizability of your results?

31. What conclusions can reasonably be drawn from your research and what cannot be said with confidence as a result of what data was used?

*Help us make this resource better. Send feedback to MSREthics@microsoft.com.*