

This glossary is in progress. Please feel free to add new entries and questions!

| Term                             | Definition                                                                                                                                                                                                                                                                                                                                                                                                                 | Discussion/Example                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Sources                                                      |
|----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| <a href="#">Context document</a> | “Interventions designed to accompany a dataset or ML model, allowing builders to communicate with users”.                                                                                                                                                                                                                                                                                                                  | Context documents are standardized documentation formats that convey information about the dataset, types of context documents include datasheets, nutrition labels, etc.                                                                                                                                                                                                                                                                                           | [3]                                                          |
| <a href="#">Data annotation</a>  | Although data annotation and labelling are often used interchangeably in ML, labelling is a subset of annotation. (See <a href="#">labelling</a> )<br><br>Data annotation refers to the process of adding information to a dataset to provide more context. For example, adding metadata.                                                                                                                                  | Annotation can include metadata about the units of measurement.                                                                                                                                                                                                                                                                                                                                                                                                     |                                                              |
| <a href="#">Data practices</a>   | “What and how data are collected, managed, used, interpreted, released, reused, deposited, curated, and so on...”                                                                                                                                                                                                                                                                                                          | Data practices are the decisions made in the collecting, interpretation, etc. of data.                                                                                                                                                                                                                                                                                                                                                                              | [2]                                                          |
| <a href="#">Extrinsic bias</a>   | Extrinsic bias refers to bias that exists within the dataset which are reflections of social, historical biases.                                                                                                                                                                                                                                                                                                           | “Extrinsic bias is concerned with a view of a biased dataset “from the outside.” The argument is that an already-biased dataset can cause even innocent software to produce a biased outcome - and may look like people saying things such as “the data made me do it.” ... If we fail to remember that a dataset is biased, then we may treat it as “fair” or “representative,” harming people who have been excluded from it.”                                    | [11]                                                         |
| <a href="#">Informed consent</a> | Informed consent is a standard ethical principle of research with human subjects that rests on the commitment that participants <ul style="list-style-type: none"> <li>- Are fully informed</li> <li>- Decide voluntarily</li> <li>- Before research is conducted.</li> </ul> Its application in online environments is complicated by the shift in technology and methods (see [4]), but the principle remains important. | In conventional human subject studies such as interviews, an IRB reviews ethics protocols and evaluates if the research is compliant with principles such as the <a href="#">proportionality principle</a> . In social media research, things get complicated. In some situations, implied consent (see [7]) can be present but must be justified. In the case of LLMS, widespread data collection without consent has prompted massive ethical and legal concerns. | A discussion of Twitter research ethics [4] is a good start. |
| <a href="#">Intrinsic bias</a>   | “The ways in which we change the data “from the inside” of data science work-processes while we are preparing the data for modeling.”<br><br>Intrinsic bias is the bias data workers introduce to the dataset.                                                                                                                                                                                                             | “Through practices of data wrangling, curation, and feature-engineering, humans make a series of decisions about how to treat their data, and those decisions may inadvertently introduce bias into the data.”                                                                                                                                                                                                                                                      | [11]                                                         |
| <a href="#">Labelling</a>        | Labelling is a specific type of                                                                                                                                                                                                                                                                                                                                                                                            | Labelling tweets on Twitter as                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                              |

|                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                          |      |
|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
|                                             | <a href="#">annotation</a> that involves assigning a predefined category to a data item.                                                                                                                                                                                                                                                                                                                                                                                                                     | 'human-generated' or 'bot-generated'.                                                                                                                                                                                                                                                                                                                    |      |
| <a href="#">PID</a> : persistent identifier | "Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. [IDs] must be persistent. It takes time and money to keep web links active, so links tend to become invalid over time. Registry services guarantee resolvability of that link into the future, at least to some degree."                                                                     | ORCID iDs are persistent identifiers for people. DOIs are persistent identifiers for journal articles, datasets, etc.                                                                                                                                                                                                                                    | [5]  |
| <a href="#">Population</a>                  | Mathematical term used to describe a group of units sharing a common trait.                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                          |      |
| <a href="#">Positionality statement</a>     | "Researcher/Practitioner Self-Disclosure: Practice should involve a disclosure of the researcher's position in the world, her or his goals, as well as the researcher's position in her or his intellectual and, to an appropriate extent, political beliefs"                                                                                                                                                                                                                                                | See [8] for examples.                                                                                                                                                                                                                                                                                                                                    | [1]  |
| <a href="#">Problem formulation</a>         | The process with which a problem is formulated and the methods we use to define and measure it define the lens with which we see the problem. "...the problems we solve with data science are never insulated from the larger process of getting data science to return actionable results..., these ends are very much an artifact of a contingent process of arriving at a successful formulation of the problem, and they cannot be easily decoupled from the process at arriving at these ends." [13:9]. | For example, O'Neil describes how proxies are used to quantify university excellence through indicators that are easily collected such as SAT scores, student-teacher ratios, and acceptance rates rather than through students' learning experience, happiness, productivity, personal fulfillment, etc. [12].<br><br>See [13] for additional examples. | [13] |
| <a href="#">Proportionality principle</a>   | In ethics, it is understood that actions have positive and negative effects simultaneously. This is called <i>double effect</i> .<br>"Applications of double effect always presuppose that some kind of proportionality condition has been satisfied. Traditional formulations of the proportionality condition require that the value of promoting the good end outweigh the disvalue of the harmful side effect."                                                                                          | In medicine, a surgeon may cause harm to a patient's skin (negative) in order to save their heart (positive). It would not be permissible for a surgeon to open up someone's chest just to get a better look or take a selfie with it because that would violate the proportionality principle.                                                          | [9]  |
| <a href="#">Provenance</a>                  | Provenance information provides a trail of history about how the data originated, how it's changed, who was involved, and more.                                                                                                                                                                                                                                                                                                                                                                              | See the following blurb from the FAIR principles...<br><br>"For others to reuse your data, they should know where the data came from ... who to cite and/or how you                                                                                                                                                                                      | [6]  |

|                             |                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                   |          |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
|                             |                                                                                                                                                                                                                                                                                                                                                             | wish to be acknowledged. Include a description of the workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else that you may have transformed or completed?" |          |
| <a href="#">Reflexivity</a> | "Questions of reflexivity ask us to consider who we should listen to and why, how to place actors' ideas in a larger field of power, questions about our own relationship to actors' theories of the world. Reflexivity asks us to approach our work with epistemological unease because we are always at risk of reproducing categories that reify power." |                                                                                                                                                                                                                                                                   | [10, 14] |

- [1] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, May 07, 2011. Association for Computing Machinery, New York, NY, USA, 675–684. <https://doi.org/10.1145/1978942.1979041>
- [2] Christine L. Borgman. 2017. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.
- [3] Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–27. <https://doi.org/10.1145/3479582>
- [4] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Soc. Media Soc.* 4, 1 (January 2018), 2056305118763366. <https://doi.org/10.1177/2056305118763366>
- [5] GO FAIR. F1: (Meta) data are assigned globally unique and persistent identifiers. *FAIR Principles*. Retrieved January 20, 2024 from <https://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/>
- [6] GO FAIR. R1.2: (Meta)data are associated with detailed provenance. *FAIR Principles*. Retrieved January 20, 2024 from <https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/>
- [7] Information and Privacy Commissioner of Ontario. Consent may be implied in some cases. *Information and Privacy Commissioner of Ontario*. Retrieved January 20, 2024 from <https://www.ipc.on.ca/part-x-cyfsa/consent-and-capacity/elements-of-consent/consent-may-be-implied-in-some-cases/>
- [8] Calvin Liang. 2021. Reflexivity, positionality, and disclosure in HCI. *Medium*. Retrieved January 20, 2024 from <https://medium.com/@caliang/reflexivity-positionality-and-disclosure-in-hci-3d95007e9916>
- [9] Alison McIntyre. 2023. Doctrine of Double Effect. In *The Stanford Encyclopedia of Philosophy* (Winter 2023), Edward N. Zalta and Uri Nodelman (eds.). Metaphysics Research Lab, Stanford University. Retrieved January 20, 2024 from <https://plato.stanford.edu/archives/win2023/entries/double-effect/>
- [10] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP (January 2022), 1–14. <https://doi.org/10.1145/3492853>
- [11] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences.

In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022. ACM, New Orleans LA USA, 1–19. <https://doi.org/10.1145/3491102.3517644>

- [12] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
- [13] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* '19)*, January 29, 2019. Association for Computing Machinery, New York, NY, USA, 39–48. <https://doi.org/10.1145/3287560.3287567>
- [14] Paige L. Sweet. 2020. Who Knows? Reflexivity in Feminist Standpoint Theory and Bourdieu. *Gend. Soc.* 34, 6 (December 2020), 922–950. <https://doi.org/10.1177/0891243220966600>