# Ethical, Legal, and Social Implications of Functional Genomics Data Generation and Downstream AI Uses: An NIH Bridge2AI Initiative Qualitative Study

Danielle M. Pacia,[1] Ian Stevens,[1] Vardit Ravitsky,[1] Jillian A. Parker,[2] Trey Ideker,[2] Timothy Clark,[3] Jean-Christophe Bélisle-Pipon[4]

1. The Hastings Center
2. University of California San Diego
3. University of Virginia
4. Simon Fraser University

The NIH Bridge2AI (B2AI) program aims to produce AI-ready datasets that adhere to the Findable, Accessible, Interoperable, Reproducible (FAIR) principles and integrate ethical considerations in collecting and preparing the data for computation.[1] The program adopts a highly multidisciplinary approach, recruiting experts in biology, standards, computer science, ethics, and team science. One data generation project within this larger program is the Functional Genomics Grand Challenge, Cell Maps for AI (CM4AI), which generates multi-modal genomic and proteomic data for integration into large-scale maps of cell structure and function for use in downstream AI/ML systems.[2] For example, these data can enable development of "visible" AI/ML models that allow for the interpretable translation of genetic inputs (e.g., mutations) into phenotypic outputs, such as drug response or disease state, via the underlying molecular pathways.

Semi-structured interviews were conducted with CM4AI stakeholders to characterize value-laden decision-making processes within a multidisciplinary approach to generating FAIR datasets, while exploring the ethical, legal, and social implications (ELSI) of functional genomics research. This qualitative research provided nuanced insights into how decisions were made during the complex process of data generation, offering a clearer understanding of their implications for future AI development.[3,4]

## Methods

A review was conducted between 2023-2024 to critically analyze the literature relating to different ethical values in medical AI development.[5] Drawing from this review, an interview guide of questions relating to the first phases of the AI lifecycle — problem identification, data generation, and data evaluation — was constructed to query those involved in CM4AI's data generation and integration activities. Between June and November 2024, authors conducted semi-structured interviews with 18 CM4AI-affiliated stakeholders including investigators, researchers, and other key NIH participants (this purposive sample represents around 85% of the total target population). Interviews were 30-60 minutes in length and transcribed using CoLoop.ai. Authors then developed a codebook using a modified constructivist grounded theory approach and used Dedoose, a qualitative analysis software to code the transcipts.[3] Using an iterative, thematic coding approach, three team members (DP, IS, JCBP) coded the interviews and conducted weekly joint coding sessions to ensure intra- and inter-coder reliability.

The interviews examined how biomedical researchers, AI developers, data scientists, and other stakeholders incorporate ELSI considerations into functional genomics research within the context of multidisciplinary team science and the broader B2AI Consortium. Themes centered on key decision points that influence and shape data generation and AI/ML model development. This study was approved by Simon Fraser University's IRB.

## Results

Through qualitative analysis, ELSI themes emerged relating to functional genomics research and data generation practices for downstream AI use, including Problem Identification and Prioritization, Data Accessibility/Usability, Data Sourcing, Transparency/Explainability, Generalizability, and Bias (See Table 1). Insights also addressed team science in generating AI-ready datasets (See Table 2).

## Discussion

While some decision-making related to the ELSI that emerged during data generation was relatively straightforward, such as ensuring that values of transparency and explainability were prioritized, other decisions were more difficult and required nuanced reasoning processes. For example, cell line selection required balancing several different ethical considerations, including informed consent, diversity and inclusion, and reliability. These difficult decisions were supported through multidisciplinary reasoning, weighing the benefits and drawbacks of selecting a particular cell line.[6] Such multidisciplinary ethical reasoning is especially important early in the AI lifecycle when decisions may have outsized downstream negative effects.

At a high-level, results point to the importance of careful multidisciplinary deliberation early in the AI lifecycle. Future scholarship should evaluate how such multidisciplinary team science approaches influence ethical deliberation during data generation. For stakeholders in functional genomics, these findings underscore the ongoing need for integrative and robust ethical frameworks that accommodate evolving AI applications and data-driven innovation.

### Funding Acknowledgment

*This research was supported through the Bridge2AI program, NIH Grant Number: 1OT2OD032742.*

## References

1. Bridge to Artificial Intelligence (Bridge2AI) | NIH Common Fund. Accessed December 25, 2024. https://commonfund.nih.gov/bridge2ai

2. Clark T, Mohan J, Schaffer L, et al. Cell Maps for Artificial Intelligence: AI-Ready Maps of Human Cell Architecture from Disease-Relevant Cell Lines. Preprint. bioRxiv. 2024;2024.05.21.589311. Published 2024 May 24. doi:10.1101/2024.05.21.589311

3. Charmaz K. Constructing Grounded Theory. Sage Publications; 2006.

4. Hoeyer K, Koch L. The ethics of functional genomics: same, same, but different? Trends in Biotechnology. 2006;24(9):387-389. doi:10.1016/j.tibtech.2006.06.011

5. Victor G, Barbu A, Bélisle-Pipon JC. Moral Values in Medical AI: A Scoping Review. Published online May 16, 2024. doi:10.21203/rs.3.rs-4391239/v1

6. Pacia DM, Ravitsky V, Hansen JN, Lundberg E, Schulz W, Bélisle-Pipon JC. Early AI Lifecycle Co-Reasoning: Ethics Through Integrated and Diverse Team Science. The American Journal of Bioethics. 2024;24(9):86-88. doi:10.1080/15265161.2024.2377106

# Table 1

| ELSI Value and Thematic Finding | Representative Quotations |
|---|---|
| **Problem Identification and Prioritization** The initial stage where the specific issue or challenge that an AI system is designed to solve is clearly defined and understood. | "It comes down to… what genes do we perturb…So on that front, we made the decision that we only going to have so much money. We have to prioritize genes of great importance. [S3_06] |
| **Data Accessibility and Usability** CM4AI consortium members explained how they may address barriers to data sharing and how to optimize FAIR principles through clear documentation and provenance. | "Yeah, we want to make [data] that's as user friendly, as well documented, and as standardly put together as possible, so that it is the smallest lift possible needed for someone to come along and use it." [S3_03] |
| **Data Sourcing** Throughout the interviews, ethical considerations in the sourcing and curation of data were noted. Researchers weighed diversity and inclusion concerns against whether consent was documented, while also considering the reliability of the cell line. | "I think we selected which cell lines very carefully. I think they picked the ones that have been used the most, that they have the most experience with. I was very interested in the diversity. I do health disparities research. I was excited that they picked one that was for certain populations, especially aggressive cancers." [S3_13] |
| **Transparency/ Explainability** The CM4AI Grand Challenge emphasized throughout data generation that AI models should be made interpretable to users. | "The goal that CM4AI is structured around, obviously, is parallel data generation for multiple modalities in the same cell types to create a more robust or as robust of a map as we can have of cell structure and cell function. Those maps would then be put into an AI or machine learning model, and those maps become the underpinning for making the model explainable. That makes it interpretable and explainable, and that removes the quote unquote black box from our models." [S3_09] |
| **Generalizability** Participants noted that the scope of generalizability was understandably limited but is nonetheless worthwhile work. | "No, it's not generalizable. I mean, I hesitate to say anything that would make it sound like if CM4AI does this in three cell lines, that would be anything other than 1% or a fraction of 1%. The big thing in my field these days is the environment and how it influences the genome…So CM4AI cell lines are very sterile. These are very highly controlled, lab living, petri dish inhabiting things which at best can tell us what happens under very tightly controlled circumstances… This is only generalizable to the extent that it can be replicated in another lab. But it is good knowledge. It's fundamental knowledge of how the genome functions." [S3_14] |
| **Bias** Interviewees stressed the need to mitigate three interconnected and overlapping forms of bias: data-driven, human, and algorithmic. . | **Data-Driven** "Of course it's a weakness that it's only one cell line. But I'm also not too worried. If we talk about basic biology, it's like a lot of the work that's being done is in cell biology, is being done with reference data sets that might be one cell line. Like the first version of the human genome was (largely) one genome. [parenthetical added to avoid a slight factual error]." [S3_08] **Human-Driven** "So the biases in CM4AI is going to be more around the decision-making. So the decision to use certain parameters to pull from certain sources is going to introduce biases in different ways along the pipeline. Some of that is what we…[A]s a statistician, I sit there and go, well, how sensitive is your arbitrary decision to the end pipeline. And so we have a number of points that we've sort of earmarked as potential spots where we can investigate the impact of biases if we, whether we get to them all or not is a whole other issue, because it just takes time, if you change something, to rerun all of this stuff." [S3_03] **Algorithmic-Driven** I mean, that's the old statement, garbage in, garbage out. If I throw random stuff at things and then it comes out with some answer and I don't take into account what the algorithm is doing, it may come out with a map that's completely bogus, which would end up wasting their time. Or, you know, right now this is all very preliminary, but you're right, later on is if this ends up at a point where it impacts a patient decision… that all of a sudden [is] scary. And that is a, that is a very valid concern and it does make one worry. But in defense of that, all the code is open source and the algorithm is there for everyone to see. [S3_04] |

## Table 2

| | |
|---|---|
| Beyond ELSI about data generation itself, key themes emerged relating to the Team Science approach employed by the B2AI Consortium. The CM4AI Grand Challenge is comprised of personnel that worked together previously; however, a significant portion had not collaborated before.<br><br>These themes included both the benefits (diverse perspectives) and challenges (coordination complexity) of cross-disciplinary collaboration, such difficulty in defining roles and responsibilities within multidisciplinary teams at the beginning of the project. | |
| **Team Science Challenges and Benefits** | "And that's the reason to have so many people on the team that have different perspectives. I mean, that's the reason for the diversity on the team, frankly." [S3_13]<br><br>"It's hugely challenging and it's given me an opportunity to kind of to work with some great people and to learn and a lot and to get involved in some pretty leading-edge stuff on AI. My background is not AI model development. And you [may] think, "Everything is now AI models." No, it isn't… You can't do anything with biomedical data without the knowledge representation. So. But transitioning into this world, it's really helpful to really introspect and talk with colleagues about the ethics too." [S3_12] |
| **Team Science Challenges** | "One of the challenges is how do we even divide up our roles, the boundaries. We haven't worked with each other before for many modules, and (Researcher's name) has worked with the data generation modules before, but not to this level of integration." [S3_07]<br><br>"I find the consortium… very complex, it's very multidimensional." [S3_02] |