Ethics of Human vs. Formal Encoding in Scientific Knowledge Representation

Munongedzi Mabhoko, Clarkson University, mabhokm@clarkson.edu

Scientific knowledge representation seeks to make human understanding machine-readable. Ontologies such as SNOMED CT and MeSH serve as formal systems that encode biomedical knowledge in a structured, computable format [1, 2]. While these tools have transformed healthcare interoperability and data analysis, they also raise ethical concerns about how knowledge is represented, who controls it, and what forms of meaning are lost in translation.

Human encoding how people describe, categorize, and interpret knowledge remains deeply tied to culture, language, and lived experience. It reflects interpretive richness and moral reasoning, capturing ambiguity and ethical context [3]. However, it also carries personal and institutional bias. Clinicians' interpretations of "normal" or "abnormal" can reflect social assumptions embedded in their training and environment.

Formal encoding, by contrast, prioritizes precision and interoperability. Systems like SNOMED CT represent clinical entities as hierarchical concepts with logical relationships. This structure enables data integration and reasoning across institutions but also imposes rigid definitions that may erase nuance [1, 4]. In MeSH, for example, categories such as gender identity and race were historically underrepresented, mirroring systemic biases in medical literature [2 , 5].

The ethical tension between human and formal encoding lies in balancing meaning with measurability. Human encoding is flexible but subjective, formal encoding is consistent but can embed hidden bias. In practice, these systems co-construct scientific reality, deciding which concepts are recognized as valid and which remain invisible. This epistemic control has consequences for research, diagnosis, and health policy [3, 4].

Ethical ontology design requires transparency about who encodes what knowledge, under what assumptions, and for whose benefit. New frameworks such as "model cards" [6] and "datasheets for datasets" [7] propose methods for documenting provenance and intent, blending interpretive accountability with computational structure. These efforts help produce ontologies that are technically sound and socially just.

The relationship between human and formal encoding can be summarized across several dimensions:

Dimension	Human Encoding	Formal Encoding
Source of Meaning	Contextual, narrative, culturally embedded	Logical, hierarchical, computable
Bias Visibility	Visible and debatable	Hidden within structure
Ethical Concern	Subjectivity, inconsistency	Rigidity, exclusion
Primary Risk	Misinterpretation of context	Erasure of lived diversity

Hybrid ethical frameworks advocate combining formal and human encoding through participatory ontology design. By engaging affected communities and domain experts, such systems allow pluralism multiple overlapping representations of reality. This approach aligns with emerging fairness standards in AI ethics, where technical representation is understood as a moral act rather than a neutral procedure [3, 6].

In conclusion, encoding is not a passive technical activity but a process that constructs the boundaries of what counts as scientific knowledge. Both human and formal approaches shape not only what we know but how we know it. Building ethically aware ontologies requires continuous reflection and documentation ensuring that the precision of machines does not overwrite the moral agency of humans [3, 7].

References

[1] Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in Health Technology and Informatics, 121, 279–290.

[2] Lowe, H. J., & Barnett, G. O. (2009). Understanding and using the Medical Subject Headings (MeSH) vocabulary to perform literature searches. JAMA, 3, 41–51.

[3] Bowker, G. C., & Star, S. L. (1999). Sorting Things Out: Classification and Its Consequences. MIT Press.

[4] Levesque, H., & Brachman, R. (2012). Ontology and Representation: Philosophical and Ethical Foundations. AAAI Spring Symposium, 23–30.

[5] Cooper, R. S., et al. (2019). Race and Genomics: Understanding and Avoiding Misuse. NEJM.

[6] Mitchell, M. et al. (2019). Model Cards for Model Reporting. FAT*, 220–229.

[7] Gebru, T. et al. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92.