The Ethics of Classifying the World: From Library Catalogues to AI

Clare Thornley[1], Marta Bustillo[2] and Christoph Schmidt Supprian[3]
[1]Clarity Research, Ireland.
[2]University College Dublin, Ireland.
[3]Trinity College Dublin, Ireland.
cthornley@clarityresearch.eu
marta.bustillo@ucd.ie
schmidc@tcd.ie

Abstract: This paper reports on an initial exploration of knowledge classification ethics: What are the important ethical issues in how we classify knowledge and what kind of cognitive, cultural and social impacts may they have? An important part of Knowledge Management is the classification and organisation of knowledge to make it findable and reveal connections in related subjects. Discussion on the ethical aspects of this issue have recently been brought to the fore in both Library and Information Studies (LIS), in terms of objections to Library classification terms, and also in AI which can classify data using data sets which themselves reflect existing injustices and bias. The ethical implications of both types of knowledge classification can be better understood when the classification ethics debate in LIS and AI are used to inform each other. Findings include that AI provides clarity on measuring adverse outcomes whilst LIS provides nuance on the potential cultural and psychological harm of inappropriate terminology and inaccurate positioning within 'worlds of knowledge'.

Keywords: Artificial Intelligence, ethics; Library and Information Studies, classification systems, terminology

1. Introduction

The inspiration for this paper is many years of discussion and teaching on the topic of cultural bias in Library Classification systems, such as the Dewey Decimal System and the Library of Congress Subject Headings. Recently this debate has extended to concerns about bias in Artificial Intelligence (AI) systems and the ethical impacts of how we classify knowledge has received wider attention. Classification is a core part of KM (Alavi and Leidner, 2001) and thus ethical issues in classification influence KM. Classification is the act of identifying the subject of information items and then assigning them to class (category) of similar items. This then places the item within a structured (usually hierarchical) system. This immediately raises questions of how 'representative' the subject headings and their structure are of the knowledge 'in the world' and who or what may be missed out, misrepresented or marginalised. A system or a data set are biased if they are not an accurate representation (sample) of what (population) they are supposed to represent. Representation is , by its nature, incomplete and imperfect but it is deemed biased if it is systematically inaccurate, particularly when it generally disfavours certain groups. Are these systems 'biased' in terms of favouring certain subjects and people and inaccurately representing others? Is the system therefore 'unfair'? The definition of bias and , in more depth 'fairness' , is discussed in some depth in AI (Mehrabi *et al.*, 2022) We argue that understanding this issue requires both an historical and future focused perspective. It needs to combine a practical focus on negative impact with respect for the seemingly more subtle harm of mis-labelling groups of people or historical events. The current debate on bias in AI does not , as yet, acknowledge and learn from previous related work in the Library and Information Studies field and *vice versa*. The potential of AI to frame problems in ways which can marginalise currently does not include cultural /spiritual harm but rather focuses on function/resources whilst the debate in Library Classification rarely offers direct evidence of material harm. Both literatures often fail to engage with the structural inequalities which created the bias they decry. The aim of this paper is to make some initial observations on how the two areas of debate and research could inform each other and improve understanding of this increasingly vital topic.

The structure of the paper is as follows. Firstly, we define the key concept of classification and its role in representing and ordering the world with a focus on the library perspective. We then describe its connection to Artificial Intelligence which is also used to categorise the world. Finally we analyse the major themes in the literature concerning bias and ethic and conclude with suggestions for further research.

1.1 What is classification?

Classification is both a thing , a classification system e.g. the Dewey Decimal System, and an activity (someone or some algorithm places items in certain categories within the system), and a set of rules or guidelines which guide either people or algorithms on the logic and methods of deciding what goes where. It also is by its nature

historical , it organised items that are published or available, but its focus is future orientated, people look for information to inform future actions.

'While cataloguing provides information on the physical and topical nature of the book (or other item), classification, through assignment of a call number (consisting of class designation and author representation), locates the item in its library setting and, ideally, in the realm of knowledge. Arranging similar things in some order according to some principle unites and controls information from various sources.'('library classification | library science |', 2013)

The question of bias in classification systems is complex. As classification systems are created and used by humans they cannot be completely 'objective 'but it is possible to make them more fit for purpose and to minimise the harm caused by inaccuracies and omissions. Classification is one method of organising ,structuring and describing the world. As such, like any method of describing , it both reflects a particular human perspective on the world and can also reinforce or undermine particular perspectives. Its purpose is mainly to enable information to be found and understood. Thus it has very real impact on research and society as it can control who finds what information and thus have a corresponding effect on their actions. Recently , for example , there has been growing awareness that the male body has been used as a model in car safety tests with the spurious assumption that this would be equally valid for women (Criado Perez, 2019). In that case a category is mis-labelled 'human' whereas in fact its content is only 'men' , this failure of accuracy causes harm.

A topic which is much discussed in the Library and Information literature is the more subtle effect of bias through its claims to represent the world and 'framing' subjects or problems in certain ways. If it excludes certain groups of people or areas of knowledge , or puts them in very small categories, or puts them in categories which are felt as inappropriate or just false (e.g. placing ethnic groups under 'history') by the people concerned it can cause harm (Moulaison Sandy and Bossaller, 2017). There is also an interest in anthropology and philosophy in terms of how the structure of language reflects the lives of the people who speak it. The title of Lakoff's book on the theory of language (Lakoff, 1987) 'Women Fire and Dangerous things' is a reference to the Aboriginal Australian Dyirbal language, in which (as interpreted by Western anthropologists) the "feminine" category includes nouns for women, water, fire, violence, and certain animals. Hence the creation and use of classification systems , both in terms of their structure and how they 'label' categories, can both reflect, influence and reinforce how we perceive ourselves and others. When this is combined with the unequal power structures and relationship between different groups of people it has the potential to further 'other' and marginalise those who are already socially and politically excluded.

1.2 What is AI?

AI has no universally agreed definition and many attempts note that we do not fully understand intelligence in humans so defining it accurately for machines is problematic. A useful definition, by the EU High-Level Expert Group on Artificial Intelligence, is given below (2018).

'Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).'

AI is concerned with classification in that one of its uses is to place data items into categories based on their characteristics and then also to 'make decisions' about what should be done with the data subjects based on an algorithm. Unlike library classification AI is often 'bottom up' classification without necessarily a pre-existing set of categories. Large data sets are analysed to 'detect' useful patterns which can 'be found' in the data. These can then be used to inform or predict future events e.g.

Which candidates should we shortlist for interview?
What kind of youth is most likely to pull a concealed weapon?

The issue of bias in AI is usually one of the data sets which are used to 'train' the system The data set may not be representative of the actual population it will inform decisions about. Data sets can also reflect accurate but unjust distribution of characteristics (e.g. more white men are senior managers, more young black men tend to

get arrested for violent crime). A problem with AI is that it often uses **descriptive** analysis to provide **normative** analysis. In these examples , for instance , it cannot judge that this distribution of outcomes is influenced by the sexist and racist assumptions (whether conscious or not) of the decision makers involved. So women are classified as 'type of person that **should not** get senior level job' because they have been 'accurately' classified as type of person who **at present does not** tend to get a senior level job.

Both AI and Library Classification systems can have a spurious aura of natural objectivity. AI can also amplify and extend its false assumptions and/or deductions in ways which are not transparent (Ali *et al.*, 2019). Additionally, as AI use spreads, it increasingly has many users who are not experts in the relevant technology and mathematics, which can make it harder to spot bias (Srinivasan and Chander, 2021).

2. Methods

2.1 Data collection

An initial literature review covering ethics and bias in Library and Information Studies and then in AI in Google Scholar was carried out. This research is in its initial stages and a more comprehensive literature review is planned. This review was supplemented through discussions with colleagues and students.

2.2 Data Analysis

The most relevant papers were selected (7) based on the abstract. The full papers were then read to pull out the main ethical themes (7) raised and whether they included a guide to practice.

3. Results

3.1 Library Classification

A review of the works on library classification schemes is heavily dominated by Northern American publications covering issues of the classification terms of the indigenous peoples with some recent discussion on trying to change the LCSH 'illegal alien' to 'undocumented immigrant' (Ford, 2020).

3.2 AI

The work on AI does not tend to cover issues of history or incongruous naming as much as classification schemes. Bias is more often discussed from a statistical perspective and normally combined with clear evidence of material negative impact on particular groups. Guidance on mitigating measures to reduce the impact of bias is more frequently included and the potential of AI to be less biased than humans (e.g. in recruitment) is sometimes mentioned.

3.3 Summary Table of themes

Source	Omission	Naming incongruent with self-assigned name	Lack of detail	Consigning to history	Inaccurate	Othering/ Defining as non-x	Material Negative impact	Review & Guidance
AI and Data								
(Ali, 2019)					Y	Y	Y	Y
(Criado Perez, 2019)	Y		Y		Y	Y	Y	Y
(Srinivasan,et al. 2021)					Y		Y	Y
Library Classification Schemes								
(Marshall, 1977)	Y	Y				Y		
(Ford, 2020)		Y						Y
(Lee et al, 2021)		Y	Y	Y	Y			
(Moulaison et al, 2017)		Y	Y	Y	Y	Y		Y

4. Discussion and Conclusions

This initial work demonstrates the potential of exploring further the relationships and mutual learning opportunities between research on the ethics of library classification and AI ethics. There are shared perspectives , in terms of concerns for the representation of traditionally disadvantaged groups, but differences in terms of the complexity of discussion regarding the nature of bias and its impact. Future work will include a more systematic literature review which will include professional guidance or standards provided for librarians or AI professionals. Our findings show that a more nuanced understanding of the nature of harm caused by classification and the role of structural power inequalities could be developed through further integration and analysis of the two literatures. It would combine cultural and spiritual marginalisation with improving understanding of material harm. This analysis should not serve as a potential distraction from efforts to correct structural inequalities but rather as a useful contribution by revealing the power of knowledge representations to both reflect and shape perceptions and experiences. A more thorough understanding of the actual nature of representation and classification, and the ethical issues it can raise will assist in making more informed and relevant decisions about how to reduce harm.

Acknowledgements

The authors would like to thank the Information and Library Management MSc students at Dublin Business School who have contributed to many interesting discussions on this topic.

References

Alavi, M. and Leidner, D. E. (2001) 'Review: Knowledge Management And Knowledge Management Systems: Conceptual Foundations And Research Issues', *Management MIS Quarterly*, 25(1), pp. 107–136.

Ali, M. *et al.* (2019) 'Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Criado Perez, C. (2019) *Invisible women : data bias in a world designed for men*. New York: Abrams.

Ford, A. (2020) 'Conscientious Cataloging | American Libraries Magazine', *American Libraries*.

High-Level Expert Group on Artificial Intelligence (2018) *A definition of AI: main capabilities and scientific disciplines*.

Lakoff, G. (1987) *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, *Language*. Chicago: The Univ. of Chicago Press.

Lee, T., Dupont, S. and Bullard, J. (2021) 'Comparing the Cataloguing of Indigenous Scholarships: First Steps and Finding', *KO KNOWLEDGE ORGANIZATION*, 48(4), pp. 298–306.

'library classification | library science |' (2013) *Britannica*.

Marshall, J. K. (1977) *On equal terms : a thesaurus for nonsexist indexing and cataloging*. New York: Neal-Schuman.

Mehrabi, N. *et al.* (2022) 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys (CSUR)*, 54(6).

Moulaison Sandy, H. and Bossaller, J. (2017) 'Providing Cognitively Just Subject Access to Indigenous Knowledge through Knowledge Organization Systems', *Cataloging & Classification Quarterly*, 55(3), pp. 129–152.

Srinivasan, R. and Chander, A. (2021) 'Biases in AI systems', *Communications of the ACM*, 64(8), pp. 44–49.