The Respective Roles of Intellectual Creativity and Automation in Representing Diversity: Human and Machine Generated Bias †

Vanda Broughton

University College London, Department of Information Studies, Gower Street, London WC1E 6BT, <v.broughton@ucl.ac.uk>

Vanda Broughton is Emeritus Professor of library and information studies at University College London. Her principal research interest is in the development of faceted classification, particularly as it affects different disciplines. She is editor of the second edition of the *Bliss Bibliographic Classification* (*BC2*), and an associate editor of the Universal Decimal Classification. In addition to the published volumes of *BC2*, she is author of several books on knowledge organization systems and numerous articles and conference papers.

Broughton, Vanda. 2019. "The Respective Roles of Intellectual Creativity and Automation in Representing Diversity: Human and Machine Generated Bias." *Knowledge Organization* 46(8): 596-606. 82 references. DOI:10.5771/0943-7444-2019-8-596.

Abstract: The paper traces the development of the discussion around ethical issues in artificial intelligence, and considers the way in which humans have affected the knowledge bases used in machine learning. The phenomenon of bias or discrimination in machine ethics is seen as inherited from humans, either through the use of biased data or through the semantics inherent in intellectually-built tools sourced by intelligent agents. The kind of biases observed in AI are compared with those identified in the field of knowledge organization, using religious adherents as an example of a community potentially marginalized by bias. A practical demonstration is given of apparent religious prejudice inherited from source material in a large database deployed widely in computational linguistics and automatic indexing. Methods to address the problem of bias are discussed, including the modelling of the moral process on neuroscientific understanding of brain function. The question is posed whether it is possible to model religious belief in a similar way, so that robots of the future may have both an ethical and a religious sense and themselves address the problem of prejudice.

Received: 18 September 2019; Revised: 12 November 2019; Accepted 14 November 2019

Keywords: machine intelligence, bias, human, artificial intelligence, data

† Presented at ISKO-UK 2019: The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization, at City University, London, UK, July 15-16, 2019.

1.0 What is artificial intelligence?

There are many and varied definitions of artificial intelligence, and various synonyms for it. Poole et al. (1998, 1) note that "the term 'artificial intelligence' is a source of much confusion," preferring to call it "computational intelligence," although it is likely that artificial intelligence is today the more widely recognised term. Other names include "machine intelligence," "synthetic intelligence" (Brachmann 2005; Gorg et al. 2014), and "augmented intelligence" (Ojala 2018; Albrecht et al. 2015; Hannay 2014)

The *Encyclopedia Britannica* (Copeland 2019) defines artificial intelligence in the following manner:

> Artificial intelligence (AI) [is] the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

The same article identifies five key aspects of intelligence, whether human or machine: learning, reasoning, problem solving, perception, and language. At the operational level, machine intelligence can take a number of forms: pattern recognition, voice recognition, image (including facial image) recognition, and machine translation. Copeland's definition (above) tends towards the narrower field of machine learning: "a form of AI that enables a system to learn from data rather than through explicit programming … "After a model has been trained, it can be used in real time to learn from data" (Hurwitz and Kirsch 2018, 4-5).

For the purposes of this paper, artificial intelligence is considered mainly within the context of information retrieval, specifically document retrieval, and the ways in which document content can be automatically identified using intelligent agents. This may involve the construction of automatic classifiers through machine learning and the way in which document content is processed.

2.0 Artificial intelligence as a complement to human activity

Artificial intelligence has impacted on many areas of human activity, in part because of the speed with which it can process information in an overloaded world, saving human effort and apparently offering a more objective way to assess and respond to a variety of situations.

Information management is only one of such uses of AI, where it may promise to solve the perennial problem of organization and retrieval in a situation where there is "too much to know." This situation has been acknowledged since the early modern period (Blair 2010), began to be addressed by mechanization in the mid-twentieth century, and in the twenty-first century prompted numerous studies of the way in which machines might automatically analyse, categorize, index, and classify documents and other information objects through a process generally referred to as automatic metadata generation or AMG (Broughton, Palfreyman, and Wilson 2008; Greenberg et al. 2005).

During that period there had also been a good deal of research into the relative roles of controlled vocabularies and automatic indexing, generally leading to the conclusion that a hybrid model offered the best balance between efficiency and effectiveness; a number of studies demonstrated that the use of a controlled vocabulary improves the performance of the tool or system (Liang et al. 2006; Cheung et al. 2005; Aula and Kaki 2005; Ko et al. 2004). Another major theme was the automatic building of classificatory structures such as ontologies, independently of humans, from text corpora or other sources. The extraction of data from text continues to be a common means of constructing semantic tools, but, as we may see below, the assumption that terms in text are value free, and mean exactly what they say, presents a danger to the usefulness and efficiency of such exercises.

As AI gains ground as an established tool for processing and decision making, particularly with respect to personal data, some questions have been raised as to the acceptability of AI in this role, the ethics of AI, and the extent to which machines can function as intelligent, and as ethical agents. Associated ideas, such as the personhood of robots, and whether they can be said to assume responsibility for their actions, have also been considered.

3.0 Ethical considerations in knowledge organization

It is now well established that the business of knowledge organization, whether that is classification, indexing, subject representation through headings, or visualization tools, brings with it some ethical concerns. Recently, attention has focussed on fake news, or controversial thinking, such as holocaust denial, and how such material should be represented, but more generally concerns are with the misrepresentation or under representation of minority groups, leading to disadvantage and disempowerment. There is now a growing body of literature on the ethical theory and philosophy of KO (Olson 1998; Szostak 2014; Mai 2010, 2013a, 2013b, 2016), and a substantial number of studies of the way in which it can discriminate on the basis of gender (Foskett 1971; Marshall 1977; Olson and Ward 1997; Olson 2007), sexual orientation (Drabinsky 2013; Fox 2016; Howard and Knowlton 2018), race and ethnicity (Duarte and Belarde-Lewis 2015; Adler and Harper 2018), political status (Lacey 2018), and religion (Broughton 2000; Broughton and Lomas 2019).

All of such bias is problematic in a world of increasing diversity, and the major players in conventional KO are seen to address some of the worst excesses. Factors which exacerbate the bias include: unequal provision either of terminology or (in a coded system such as a classification) unequal distribution of notation; failure to name at all certain groups or perspectives; and language which has a strong flavour of one particular favoured perspective or culture. Where culture is a powerful element, as in religions, language is a specific problem.

4.0 Ethical considerations in artificial intelligence

There has been substantial research into the phenomenon of machine ethics, that is the potential ethical or moral behaviour of intelligent agents. It should be carefully differentiated from computer ethics which is concerned with the behaviour of humans in the context of computing and information technology, and with roboethics which refers to ethical behaviour of humans in the design and construction of intelligent machines, and in human-machine interaction. Although there are some twentieth-century discussions of the possibility of moral—or immoral—actions of machines, the field really begins with the 2005 *AAAI Symposium on Machine Ethics*, where the problem is clearly stated, and named by Anderson et al. (2005):

> Past research concerning the relationship between technology and ethics has largely focused on responsible and irresponsible use of technology by human beings, with a few people being interested in how

> human beings ought to treat machines. In all cases, only human beings have engaged in ethical reasoning. We believe that the time has come for adding an ethical dimension to at least some machines. Recognition of the ethical ramifications of behavior involving machines as well as recent and potential developments in machine autonomy necessitate this. We explore this dimension through investigation of what has been called machine ethics.

In Anderson et al.'s paper, they consider the implementation of two systems of machine ethics, based on philosophical principles as displayed in the work of W. D. Ross (theory of prima facie duties), and Jeremy Bentham (Utilitarianism), both of which can be expressed as a series of rules. Utilitarian ethics and the basis of its decision-making is of particular interest, since in the form of Bentham's *Felicific calculus,* or *Calculus of pleasures* (1789), it was designed to be computable, and indeed, one of the themes of the Symposium was the computability of ethics. At the time of Anderson et al.'s research, the likely guarantee of "good" machine ethics was the imposition of better and more considered rules for the machine's operation, derived from traditional systems of ethics and the practice of professional ethicists. As they say in a subsequent paper (2007, 25):

> Ensuring that a machine with an ethical component can function autonomously in the world remains a challenge to researchers in artificial intelligence who must further investigate the representation and determination of ethical principles, the incorporation of these ethical principles into a system's decision procedure, ethical decision making with incomplete and uncertain knowledge, the explanation for decisions made using ethical principles, and the evaluation of systems that act based upon ethical principles.

4.1 Where machine ethics falls short: bias in intelligent agents

The general assessment of machine information processing and machine decision-making has been that it may avoid the subjectivity associated with humans. In practice this has turned out not to be the case, since, despite the emphasis on machine independence in artificial intelligence, intelligent agents are not created spontaneously, but require some degree of human participation, and no system of machine learning can avoid the use of information which has been at some stage processed by humans. The problem affects equally machine learning where the agent has learned from data or a prepared model or training set, or in the case of knowledge organization systems, where human-constructed vocabularies or ontologies have been sourced by the agent. Additionally, human intervention often supports the machine-learning process through iteration with the "teacher," usually through a technique of query-by-example accompanied by feedback to the machine.

Too often the result of this human input is that the machine inherits the prejudices of the human, so that the bias is hard-wired to the machine (Crawford 2016; Kirchner et al. 2016; Sears 2018; Kochi 2018). The World Economic Forum (2018, 3) stated:

> Designed and used well, machine learning systems can help to eliminate the kind of human bias in decision-making that society has been working hard to stamp out. However, it is also possible for machine learning systems to reinforce systemic bias and discrimination and prevent dignity assurance.

The likelihood of such bias has considerable implications for human rights, for the proper management of social diversity, and for the fair treatment of diverse groups in society. Such inequity is a long-standing problem in conventional information management and has been addressed at length in the research literature of knowledge organization in particular. It seems, however, especially insidious in the machine intelligence context, perhaps because of the expectation that higher levels of neutrality and objectivity apply.

4.2 Bias and discrimination derived from data

Much of the literature in this area is centred on machine decision-making based on demographic data, and the concern arises from a human rights perspective where some groups are disadvantaged or marginalized by the way in which the data is set up (Smith, , Patil and Muñoz 2016; Obama White House 2016; World Economic Forum 2018). Generally in these cases, the data is factual and the decision-making is based on a combination of values in different categories, and the identification and recognition of patterns embedded in data, especially latent associations between one group of attributes and another.

Particularly prominent in the discussion of bias in AI is gender discrimination, also a feature of early research into bias in KO. A recent major study by Criado-Perez (2019) reveals that data itself is often biased, because the sample is in some way flawed. Criado-Perez's principal concern is with gender imbalance, and it is clear that a female perspective is often omitted, because the data is derived from studies that dealt only with males. Criado-Perez provides examples of where, for example, diagnostic thresholds based on biomarkers are inaccurate for women, because average figures are based on predominantly male data (as in Khamis et al. 2016), since the inclusion of female data

may be as low as 14% of studies in some fields (Pinnow et al. 2014). This may explain many examples of bias detected in machine intelligence where the agents have been trained on datasets that lack comprehensiveness in one or more respects.

4.3 Bias and discrimination derived from semantics

A much-cited paper by Caliskan et al. (2017) establishes that not only is any incompleteness or skew in the data sample passed on to intelligent systems, but that semantics is also an inheritable factor. Using measurable associations between pairs of words, Caliskan builds on the work of some prior studies investigating human-like biases in textual corpora, particularly that of Greenwald (1998), who studied "biases that they consider nearly universal in humans and about which there is no social concern" (Caliskan, 183). Clear associations between "flowers" and "pleasant," and "insects" and "unpleasant," were replicated by Caliskan's team, as were similar links between "weapon" (unpleasant) and "musical instrument" (pleasant), proving the soundness of the methodology. Caliskan et al. were also able to replicate more socially significant connections between European American names and African American names with pleasantness and unpleasantness respectively, a phenomenon confirmed in practice by Bertrand and Mullainathan (2004) who tested employers' response to job applications varying only in the attached European or African sounding names.

Comparable work was also replicated in the area of gender, associating female names with "family" as opposed to "career," when compared with male names (Nosek et al. 2002a), and the correlation of women with the arts, rather than mathematics, or with the sciences (Nosek et al. 2002b).

Studies of inherent discrimination based on religion are much less frequent, perhaps because religious affiliation is much less immediately obvious than gender or ethnicity. However, Binns (2018, 1) places it on a level with gender and race as a potential factor for discrimination, using the example of disparate treatment of nationals from Muslim-majority countries because of a perceived association of Islam with terrorism (2018, 4). Since such examples of religious prejudice are not uncommon and the problem is one with high public awareness, it is surprising that, to date, there is little or no research into bias associated with religious affiliation.

The existence of such semantic bias has considerable implications for information retrieval because so many automatic classifiers build structures on the back of text corpora on the assumption that these present a neutral and objective picture of the world. The existence of these biases seem to be clearly acknowledged in the world of corpus linguistics, but did not seem to be taken into account at all in the field of automatic classification or term extraction tools, perhaps because work on these originated in the sciences rather than the social sciences and humanities. Automated lexicography is now a very well-established methodology for building such semantic tools, whether the lexical data is sourced from other lexical tools such as dictionaries or extracted from text corpora, and the problem of inherited bias could consequently be a serious impediment to both effective retrieval and ethical practice.

5.0 Inherited semantic bias in religious terminology: the example of WordNet

WordNet is a vocabulary database, maintained at Princeton University, and used extensively as semantic content for all kinds of automatic indexing and classification tools. It defines itself in the following way (Princeton University 2010):

> WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser … WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet is displayed in a thesaurus-like format, although the tags it uses are not the conventional ones of information science, but rather offer a more analytical and nuanced range of inter-term relations. Nevertheless it approximates to the standard tags through its use of the categories hypernym (= broader term, superordinate class), hyponym (= narrower term, subordinate class), and synsets (equivalence relationships). Other relationships include types (= narrower term generic), instances (= narrower term instantive), and meronymy (= narrower term partitive). Opening up a "sister term" reveals terms in the same array, comparable with some kinds of related, or associative terms. This suite of relationships provides the vocabulary with a robust logical structure, and verbs as well as nouns are thus organized into hierarchies. Unlike most controlled vocabularies, WordNet also includes adjectives and adverbs in its database. However there is quite limited use of associative term type links, so the navigation tends to be, on the whole, hierarchical.

WordNet is a good example of a resource that has inherited content. Although initially it was intellectually constructed, it draws on older sources such as thesauri (Barocas et al. 2018) that themselves may contain bias, and because of its widespread use in computational linguistics it passes on that bias.

In some applications, researchers repurpose an existing scheme of classification to define the target variable rather than creating one from scratch. For example, an object recognition system can be created by training a classifier on ImageNet, a database of images organized in a hierarchy of concepts. ImageNet's hierarchy comes from Wordnet, a database of words, categories, and the relationships among them. Wordnet's authors in turn imported the word lists from a number of older sources, such as thesauri. As a result, WordNet (and ImageNet) categories contain numerous outmoded words and associations, such as occupations that no longer exist and stereotyped gender associations.

Even a cursory examination of WordNet's religious categories reveals some very evident examples of bias. As with many humanities and social science disciplines, particularly those where there is a strong cultural dimension, language is a source of some problematic classes and linguistic expressions. Similarly, the precise analytical structure of WordNet, based on linguistics principles, while it is highly suitable for the sciences, does not always serve the rather messier humanistic domains nearly as well. Accurate and comprehensive category structure is not necessarily to be found, and in many cases the arrays are incomplete.

If we consider the standard criticisms of biased religion classes in standard bibliographic classifications, many of the same shortcomings are evident in WordNet. For example, if we look at the hierarchical display of hyponyms (subordinate classes) under Religion (disregarding the annotations and further levels of hierarchy) we find:

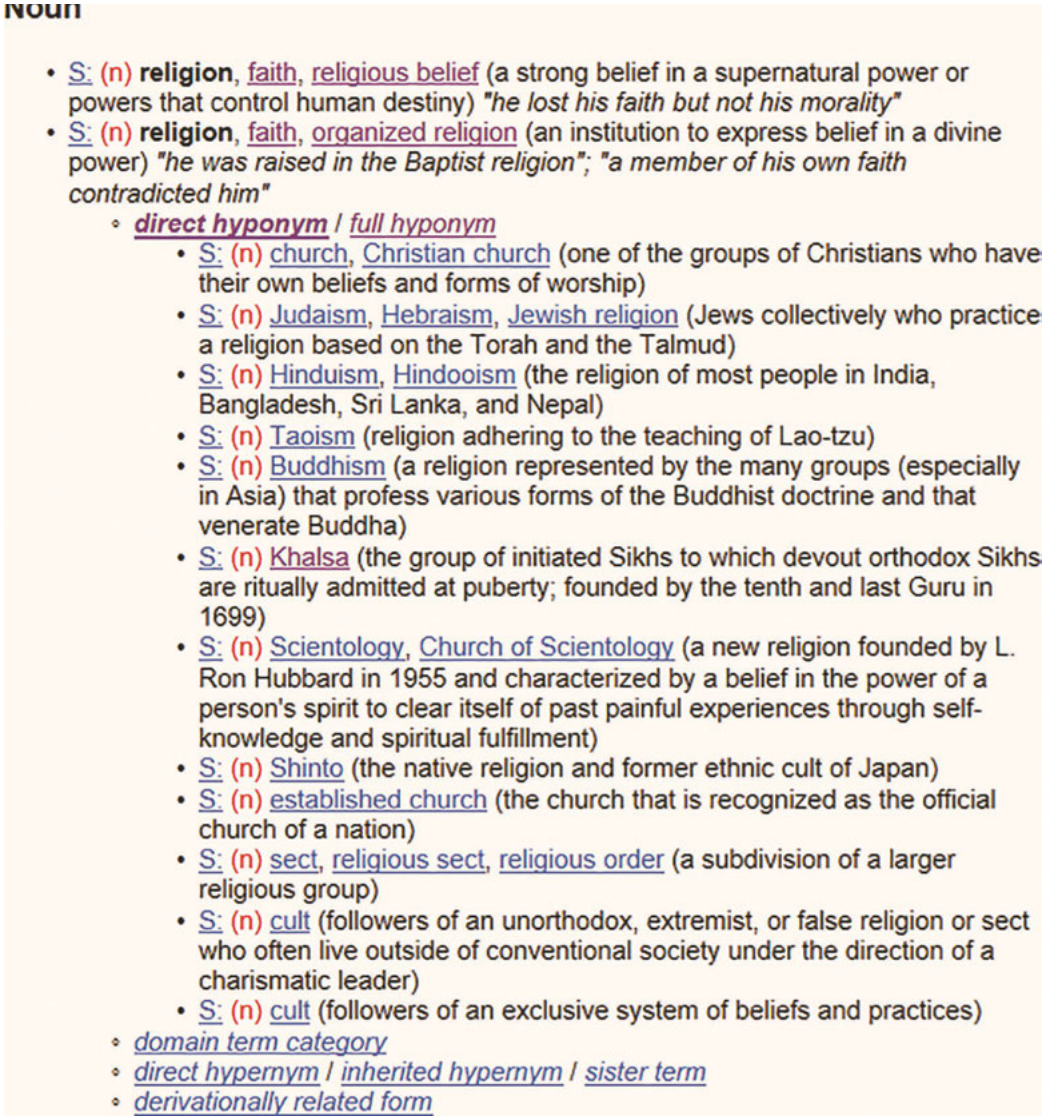

Figure 1. Entry for "religion" in WordNet.

When compared with the standard "big twelve" religions acknowledged by most sources (for example, Hinnells 2017; Boyett 2016), WordNet fails to mention Baha'i, Confucianism, Jainism, Sikhism (other than through its subset Khalsa), Zoroastrianism, and, amazingly, Islam. Expanding the list to include "full hyponyms" expands the hierarchy and brings in various Christian denominations, movements within Judaism and Buddhism, and under sects, Anglican High Church, Sunni and Shi'a Islam, the Society of Friends or Quakers, Jainism and Hare Krishna. Perhaps surprisingly, Scientology appears, but not the Mormon Church.

Needless to say, the omissions, odd associations and peculiar language (Hindooism looks very antiquated and mildly offensive) would be unacceptable in a modern thesaurus or bibliographic classification. Many of the definitions and verbal qualifications of entries exhibit some odd if not doubtful attitudes, as in the definition: "Hindooism (a body of religious and philosophical beliefs and cultural practices native to India and based on a caste system." There are differing schools of thought about the caste system, and whether it arises from socioeconomic rather than religious forces, and this is a very contentious statement. Similarly, Paganism (synonyms: pagan religion, heathenism) is defined as "any of various religions other than Christianity or Judaism or Islamism" which is certainly inaccurate in respect of modern pagans, and potentially offensive to followers of the non-monotheistic faiths. Perhaps the worse sufferer is Islam, which is provided with the synonyms Muslimism, Mohammedanism, Muhammadanism, and Islamism. The Oxford English Dictionary says of Mohammedanism that "its use is now widely seen as depreciatory or offensive," and none of these terms feel very appropriate or polite. It seems likely that this rather uncomfortable content has been imported from much older dictionaries without review or amendment.

Along with such archaic uses of language, there is also a leaning towards a strongly Christian-flavoured understanding of religious terminology, as opposed to a more multi-faith approach; examples of this can be seen in the table below. The terms have been chosen as relatively neutral ones which occur in a variety of religions, but in defining the terms or providing synonyms WordNet imposes a broadly Christian interpretation (see figure 2).

In fairness to WordNet, it does contain many religion specific terms (bhakti, Gemara, hajj, lama, menorah, nirvana, shaman, Sufi, synagogue, etc.), but because of the mainly hierarchical structure these are not easily accessed through the parent religion. This positive feature needs also to be set against the general Christian tenor of the vocabulary and the evident tendency to view religion through a Christian lens.

There are also some straightforward factual inaccuracies in WordNet, for instance the qualifier of Hinduism "(the religion of most people in India, Bangladesh, Sri Lanka, and Nepal)," whereas the dominant religion in Bangladesh is Sunni Islam, followed by 83.4% of the population (Sawe 2019).

These significant shortcomings demonstrate a very considerable bias and a disregard for fairness and sensitivity towards minority groups. The gravity of bias in WordNet is considerably magnified by its widespread use as a lexical source for automatic classifiers, which implies that the prejudices will indeed have been inherited and reinforced by a great number of other intelligent agents.

Source term	Synonyms
altar	Communion table, Lord's table
baptism	a Christian sacrament signifying spiritual cleansing
bless	make the sign of the Cross over someone
festival	religious festival, church festival
monk	Brother, Carthusian, Trappist, Cistercian
preaching	an address of a religious nature usually delivered during a church service
scripture	Bible, Christian Bible, Holy Writ, Word (the sacred writings of the Christian religions)
service	church service, prayer meeting, chapel service, vesper
sin	mark of Cain

Figure 2. Synonyms in WordNet.

6.0 What solutions exist to the problem of bias?

As is the case in library and information science, where the interests and priorities of the user community demand a privileging of those interests, bias is not always regarded as a bad thing. It is generally agreed that the very fact of a specific perspective unwittingly and unavoidably generates bias towards the favoured group (such as classification schemes for libraries with specific religious affiliations). Given the importance of meeting user expectations and the needs of the user community, bias can be seen as an ethically-neutral phenomenon.

In other cases the investigation of bias is simply a part of the scholarly study of society and the legitimate search for patterns and trends in human cultures. For example, a paper by Kozlowski et al. (2019, 38) shows how the machine analytical technique of word embedding can help to reveal changes in social attitudes over time and historic changes in word meanings. In different situations, the identification of bias may be the preliminary to addressing it in a social and political context, and is a useful tool in highlighting social inequalities.

In a wider context however, bias should be energetically tackled if the system is not to appear as the tool of a particular cultural, political, or disciplinary community. Bias inherent in data is generally regarded as undesirable and has generated an area of research activity under the general heading of machine-learning fairness. Barocas, Hardt, and Narayanan (2018) provide a broadly-based survey of a number of problems and potential solutions, based on statistical adjustment. The book "offers a critical take on current practice of machine learning as well as proposed technical fixes for achieving fairness."

Mancuhan and Clifton (2014) also propose a statistical solution to bias in data used for automatic financial decision-making, employing Bayesian techniques to identify and automatically correct bias. This is incidentally one of the few papers to reference religion as an attribute subject to bias, although the authors do not go on to include it in their study.

6.1 A moral and religious solution

As with every other area of human life, machine intelligence has impacted religious communities, apart from the general philosophical questions of whether robots can act as moral agents. A number of applications exist which aim to support religious practice, such as the Roman Catholic Confession app (Rau 2011) and Muslim Pro which can tell you prayer times and the direction of Mecca in your own town or village (Muslim Pro 2019), and attempts have already been made to use robots in ritual. Most of the literature here is in popular journals and the press, so it may be difficult to assess how serious these efforts are. We learn of a Christian robot priest in Wittenberg which radiates light from its hands and pronounces blessings in five languages as part of an exhibition to celebrate 500 years since the invention of printing technology, instrumental in the Reformation and the rise of Protestantism (Sherwood 2017). Other cases include a robot Buddhist monk in China (Tatlow 2016) which reads scripture and can answer questions, and another in Japan (Field 2017) which can "chant prayers and tap drums as part of a funeral ceremony."

There is also a literature in the overlap between religious philosophy and AI that considers the nature of the relationships between intelligent agents, humans and the person of God, typically whether the creation of intelligent agents in some sense mirrors the creation of humans (Herzfeld 2003), and if the possibilities of transhumanism through the technological alteration of species are realizable (Dumsday 2017). Vidal (2007, 930) makes a comparison between man's interaction with artificial beings and his interactions with the gods, asking whether the similarities are not caused by uncertainty:

> But it is also true that where interaction is supposed to exist between the gods and their worshippers, there always remains a strong element of uncertainty which cannot easily be dismissed concerning the exact ontological nature of the hybrid arrangement by which the divinity's presence is made manifest. It is precisely the same sort of ontological uncertainty that one finds expressed in the field of robotics. And this is also why robots both fascinate and worry the general public.

6.2 The moral and religious life of machines

A pressing question is whether a real sense of moral responsibility can be developed in intelligent agents, or, more fancifully perhaps, a proper religious sense. In human beings, it may seem obvious that ethical decisions differ in some significant respect from other kinds of decisions, and that intellectual reasoning is subordinate to, or at least strongly influenced by, emotional intelligence. As Liao (2016) says:

> Central area of intellectual inquiry across different disciplines involves understanding the nature, practice, and reliability of moral judgments. For instance, an issue of perennial interest concerns what moral judgments are and how moral judgments differ from nonmoral judgments. Moral judgments such as "Torture is wrong" seem different from nonmoral judgments such as "Water is wet." But how do moral judgments differ from nonmoral, but normative judg-

ments such as "The time on the clock is wrong" or "Talking with one's mouth full is wrong"?

However, work in neuroscience has questioned whether this distinction between cognitive and emotional aspects of moral judgements is valid, suggesting instead that all such decisions depend on reasoning through complex calculation rather than a response to stimulus (Woodward 2016). Quartz (2009, 214) states:

> Perhaps the most surprising finding to date is that core emotional structures, including the midbrain dopamine system and insula, decompose uncertain choice contexts along the statistical dimensions that are the cornerstone of FDT [Financial decision theory] … recent findings suggest that the encoding of value in midbrain dopamine areas might underlie an early implicit encoding that is signaled to orbitofrontal cortex, where it guides choice.

Recent studies have shown that it is possible to identify precisely areas of the human brain responsible for social and moral behaviour, and to link deficits in brain function to immoral, or amoral, behaviour (Damasio 1994; Shoemaker 2012). Shoemaker (2012, 807) states:

> The basic limbic emotions are those present in all mammals emanating from phylogenetically analogous brain structures collectively called the limbic system … These are fear, anger, disgust, sadness, and happiness; they function chiefly to promote the survival of the individual. The moral emotions, the product of the social brain network, arise later in development and evolution (Adolphs 2003). They are guilt, shame, embarrassment, jealousy, pride, and altruism; they function to regulate social behaviors, often in the long-term interest of a social group rather than the short-term interest of the individual person (Adolphs 2003).

Such work has considerable implications for the development of ethically responsible machines, since if the nature of decision-making can be made explicit and the process modelled, then it is, at least theoretically, possible to replicate this process in machine decision-making.

A further and more difficult question is whether it is possible to inculcate religious sensibility in machines by a similar methodology. In the speculative *Age of Spiritual Machines* (1999, 6) Kurzweil, technologist and futurist, suggests that in the distant future this will happen spontaneously as the result of technological evolution:

> Even if we limit our discussion to computers that are not directly derived from a particular human brain, they will increasingly appear to have their own personalities, evidencing reactions that we can only label as emotions and articulating their own goals andpurposes. They will appear to have their own free will. They will claim to have spiritual experiences. And people—those still using carbon-based neurons or otherwise—will believe them.'

Some specific studies of the relationship between AI and religion include William Sims Bainbridge's *God From the Machine* (2006), which investigates the question of whether religious activity might occur spontaneously in machine learning. Bainbridge and Stark (1987) proposed a general theoretical framework for the scientific study of religion, which included, among many other propositions, reasons for the emergence of cults and for cult affiliation; individuals characterized by high levels of education and social isolation, it is suggested, are more likely to participate in cults, a theory supported in part by practical testing (Bader and Damaris 1996). In 1995 Bainbridge employed the technique of neural network modelling to test his theory, and to identify the reasons for human religious behaviour. The idea of religion, however, is not part of the data supplied (484): "While not denying the possibility of God's existence, our theory attempts to explain human religious behavior without assuming the truth of religion. Therefore, there is no axiom asserting the existence of the supernatural." The program attempts to formalise communication and models a community who seek exchanges with each other in search of the basics of existence and of beneficial exchanges leading to personal rewards (486 emphasis original):

> In the scenario accompanying the program, they are called *energy, water, food, oxygen*, and *life*. Some people are producers of one or another of the first four rewards, and the simulation models the development of a little economy based on exchange of these consumable rewards. But none of the 24 people can provide each other with eternal life.

Bainbridge describes how, in the attempt to find suitable exchange partners for these more intangible rewards, the concept of the supernatural, within the context of a folk religion, may emerge spontaneously (492):

> In the subculture-evolution model, an intensely interacting group of individuals commits itself to the attainment of rewards, some of which are very difficult or even impossible to obtain. As they exchange rewards among themselves, they also exchange explanations about how to get other rewards, and in

> the attempt to satisfy each other, they magnify slightly their positive evaluation of explanations. Those explanations that can be evaluated empirically will be rejected, leaving the nonempirical (supernatural) explanations that cannot readily be evaluated. Faith will spiral upward, and the group will create a folk religion through a series of thousands of tiny communication steps.

In a more developed version of the methodology, Bainbridge (2006) concludes that this is a natural, and to some extent inevitable, process that he is able to model in various ways. He posits that it is quite feasible that, primed with appropriate theological information, machines might also "communicate ideas as if they were exchange partners engaged in theological discussion with each other" (137), and that "once separated to some degree from external control, the evolving cult develops … the end point of successful cult evolution is a novel religious culture (139)."

7.0 Conclusion

The phenomenon of bias is found to be widespread in machine intelligence, both in data per se and in semantic content derived from text corpora. Most of the studies of machine bias have focussed on demographics such as gender, race, and occasionally, social class. Although it is mentioned in a few papers as another potential focus for bias, religious affiliation has not been investigated in the same way, despite the obvious existence of religious prejudice in society. However, examination of the well-established and influential resource WordNet shows high levels of bias in its structural associations and use of language, almost certainly as a result of inheritance from vocabularies used in its original construction. Because of its widespread use in the creation of search and discovery tools, particularly automatic classifiers, WordNet is likely to have passed on these prejudices.

Addressing the problems of bias has mainly concentrated on technical solutions, but recent research in modelling the ethical behaviour of humans suggests that intelligent agents may be able to develop a sense of fairness and moral responsibility independently of humans, and one not assuming pre-programming with specific rules. Several writers speculate that, in time, a sense of the religious could also emerge, and can provide a detailed demonstration of how this might happen using similar neural network methodologies. In time, robots themselves might be equipped to deal with the phenomenon of religious prejudice.

References

Adler, Melissa and Lindsey M. Harper. 2018. "Race and Ethnicity in Classification Systems: Teaching Knowledge Organization from a Social Justice Perspective." *Library Trends* 67, no. 1: 52-73.

Adolphs, Ralph. 2003. "Cognitive Neuroscience of Human Social Behavior." *Nature Neuroscience Reviews* 4: 165-78.

Albrecht, Stefano. V., Andre M. S. Barreto, Darius Braziunas, David L. Buckeridge, Heriberto Cuayáhuitl, Nina Dethlefs et al. 2015. "Reports of the AAAI 2014 Conference Workshops." *AI Magazine* 36: 87-98. doi:10.1609/aimag.v36i1.2575

Anderson, Michael and Susan Leigh Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4: 15-26.

Anderson, Michael, Susan Leigh Anderson and Chris Armen. 2005. "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories." In *Machine Ethics: Papers from the AAAI Fall Symposium,* ed. Michael Anderson, Susan Leigh Anderson and Chris Armen. Menlo Park, CA: AAAI Press, 1-7. https://pdfs.semanticscholar.org/f36b/82a0ee77c9a87f3010f9d0335270d1a24687.pdf?_ga=2.66400620.357820807.1574389742-1902475365.1566693535

Aula, Anne and Mika Kaki. 2005. "Findex: Improving Search Result Use Through Automatic Filtering Categories." *Interacting with Computers* 17: 187-206.

Bader, Chris and Alfred Demaris. 1996. "A Test of the Stark-Bainbridge Theory of Affiliation with Cults and Sects." *Journal for the Scientific Study of Religion* 35: 285-303.

Bainbridge, William Sims. 1995. "Neural Network Models of Religious Belief." *Sociological Perspectives* 38: 483-95.

Bainbridge, William Sims. 2006. *God from the Machine: Artificial Intelligence Models of Religious Cognition.* Lanham, MD: Altmira.

Barocas, Solon, Moritz Hardt and Arvind Narayanan. 2018. "Fairness and Machine Learning: Limitations and Opportunities." http://www.fairmlbook.org

Bentham, Jeremy. 1789. *An Introduction to the Principles of Morals and Legislation.* Oxford: Clarendon.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal?" *American Economic Review* 94: 991-1013.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy." *Journal of Machine Learning Research* 81: 1-11.

Blair, Ann. 2010. *Too Much to Know.* New Haven, CT: Yale University Press.

Boyett, Jason. 2016. *12 Major World Religions: The Beliefs, Rituals, and Traditions of Humanity's Most Influential Faiths.* Berkeley, CA: Zephyros.

Brachman, Ron. 2005." Getting Back to 'the Very Idea'." *AI Magazine* 26, no. 4: 48-50.

Broughton, Vanda. 2000. "A New Classification for the Literature of Religion." *International Cataloguing and Bibliographic Control* 4: 2-4.

Broughton, Vanda, Malcolm Palfreyman and Andrew Wilson. 2008. "Automatic Metadata Generation for Resource Discovery." http://www.jisc.ac.uk/media/documents/programmes/resourcediscovery/metgenreport_final_v5.doc

Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356, no. 6334: 183-6. doi:10.1126/science.aal4230

Cheung, C. F., W. B. Lee and Y. Wang. 2005. "A Multi-Facet Taxonomy System with Applications in Unstructured Knowledge Management." *Journal of Knowledge Management* 9, no. 6: 76-91.

Copeland, B. J. 2019. "Artificial Intelligence." *Encyclopaedia Britannica.* Accessed April 14.

Crawford, Kate. 2016. "Artificial Intelligence's White Guy Problem." New York Times. 25 June.

Criado-Perez, Caroline. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men.* London: Vintage Digital.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason and the Human Brain.* New York: Grosset/Putnam.

Drabinski, Emily. 2013. "Queering the Catalog: Queer Theory and the Politics of Correction." *Library Quarterly* 83: 94.

Duarte, Marisa Elena and Miranda Belarde-Lewis. 2015. "Imagining: Creating Spaces for Indigenous Ontologies." *Cataloging and Classification Quarterly* 53, nos. 5/6: 677-702.

Dumsday, Travis. 2017. "Transhumanism, Theological Anthropology, and Modern Biological Taxonomy." *Zygon: Journal of Religion and Science.* 52: 601-22.

Field, Matthew. 2017. "This Japanese Robot Can Host Low-Cost Buddhist Funerals." *Daily Telegraph,* August 24.

Foskett, A. C. 1971. "Misogynists All: A Study in Critical Classification." *Library Resources and Technical Services* 15: 117-21.

Fox, Melodie J. 2016. "Legal Discourse's Epistemic Interplay with Sex and Gender Classification in the Dewey Decimal Classification System." *Library Trends* 64, no. 4: 687-713.

Görg, Carsten, Zhicheng Liu and John Stasko. 2014. "Reflections on the Evolution of the Jigsaw Visual Analytics System." *Information Visualization* 13: 336-45.

Greenberg, Jane, Kristina Spurgin and Abe Crystal. 2005. "Final Report for the AMeGA (Automatic Metadata Generation Applications Project)." http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Greenwald, Anthony G., Debbie E. McGhee and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74: 1464-80.

Hannay, Timo. 2014. "The Digital Academy and Augmented Intelligence." *Information Today* 31: 25.

Herzfeld, Noreen. 2003. "Creating in Our Own Image: Artificial Intelligence and the Image of God." *Zygon* 37: 303-16.

Hinnells, John. 2017. *A New Handbook of Living Religions.* Wiley Online. doi:10.1002/9781405166614

Howard, Sara A. and Steven A. Knowlton. 2018. "Browsing Through Bias: The Library of Congress Classification and Subject Headings for African American studies and LGBTQIA Studies." *Library Trends* 67, no. 1: 74-88.

Hurwitz, Judith and Daniel Kirsch. 2018. *Machine Learning for Dummies.* Hoboken, NJ: Wiley.

IBM. 2019. "Data Science and Machine Learning." https://www.ibm.com/analytics/machine-learning

Khamis, Ramzi Y., Tareq Ammari and Ghada W. Mikhail. 2016. "Gender Differences in Coronary Heart Disease." *Heart* 102: 1142-9.

Kirchner, Julia, Surya Angwin, Jeff Mattu and Lauren Larson. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks." ProPublica (blog), 23 May. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ko, Youngjoong, Jinwoo Park and Jungyun Seo. 2004. "Improving Text Categorization Using the Importance of Sentences." *Information Processing and Management* 40: 65-79.

Kochi, Erica. 2018. "AI is Already Learning How to Discriminate." Quartz (blog), March 15. https://qz.com/author/erica-kochi/

Kozlowski, Austin C., Matt Taddy and James A. Evans. 2019. "The Geometry of Culture: Analyzing Meaning through Word Embeddings." *American Sociological Review* 84, no. 5: 905-49. doi:10.1177/0003122419877135

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence.* New York: Penguin.

Lacey, Eve. 2018. "Aliens in the Library: The Classification of Migration." *Knowledge Organization* 45: 358-79.

Liang, Chun-Yan, Li Guo, Zhao-Jie Xia, Feng-Guang Nie, Xiao-Xia Li, Liang Su and Zhang-Yuan Yang. 2006. "Dictionary-Based Text Categorization of Chemical Web Pages." *Information Processing and Management* 42: 1017-29.

Liao, S. Matthew. 2016. *Moral Brains: The Neuroscience of Morality.* Oxford: Oxford University Press.

Mai, Jens-Erik. 2010 "Classification in a Social World: Bias and Trust." *Journal of Documentation* 66: 627-42.

Mai, Jens-Erik. 2013a. "Ethics, Values, and Morality in Contemporary Library Classifications." *Knowledge Organization* 40: 242-53.

Mai, Jens-Erik. 2013b. "Ethics and Epistemology of Classification." PowerPoint slides for presentation at II Congresso Brasileiro em Organização e Representação do Conhecimento 29 de maio de 2013. http://jenserikmai.info/Papers/2013_EEofClass.pdf

Mai, Jens-Erik. 2016. "Marginalization and Exclusion: Unraveling Systemic Bias in Classification." *Knowledge Organization* 43: 324-30.

Mancuhan, Koray and Chris Clifton. 2014. "Combating Discrimination Using Bayesian Networks." *Artificial Intelligence Law* 22: 211-38.

Marshall, Joan K. 1977. *On Equal Terms: A Thesaurus for Non-Sexist Indexing and Cataloging*. New York: Neal Schuman.

Muslim Pro. 2019. "The Most Popular Muslim App!" www.muslimpro.com

Nosek, Brian A., Mahzarin R. Banaji and Anthony G. Greenwald. 2002a. "Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Website." *Group Dynamics* 6: 101-15.

Nosek, Brian A., Mahzarin R. Banaji and Anthony G. Greenwald. 2002b. "Math = Male, Me = Female, therefore Math ^= Me." *Journal of Personality and Social Psychology* 83: 44-59.

Obama Whitehouse. Executive Office of the President. 2016. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Ojala, Marydee. 2018. "Digital Ethics in the STM Publishing World." *Information Today* 35: 10-11.

Olson, Hope A. 1998. "Mapping Beyond Dewey's Boundaries: Constructing Classificatory Space for Marginalized Knowledge Domains." *Library Trends* 47: 233-54.

Olson, Hope A. 2002. The Power to Name: Locating the Limits of Subject Representation in Libraries. Dordrecht: Kluwer.

Olson, Hope A. 2007. "How We Construct Subjects: A Feminist Analysis." *Library Trends* 56, no. 2: 509-41.

Olson, Hope A. and D. B. Ward. 1997. "Feminist Locales in Dewey's Landscape: Mapping a Marginalized Knowledge Domain." In *Knowledge Organization for Information Retrieval: Proceedings of the Sixth International Study Conference on Information Research*. Hague: International Federation for Information and Documentation, 129-33.

Pinnow, Ellen, Naomi Herz, Nilsa Loyo-Berrios and Michelle Tarver. 2014 "Enrollment and Monitoring of Women in Post-Approval Studies for Medical Devices Mandated by the Food and Drug Administration." *Journal of Women's Health* 23, no. 3. doi:10.1089/jwh.2013.4343

Poole, David, Alan Mackworth and Randy Goebel. 1998. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press.

Princeton University. 2010. "About WordNet." WordNet. Princeton University. https://wordnet.princeton.edu/

Quartz, Steven R. 2009. "Reason, Emotion and Decision-Making: Risk and Reward Computation with Feeling." *Trends in Cognitive Sciences* 13: 209-15.

Rau, Andy. 2011. "Should we use a Confession App?" *Think Christian (blog),* February 24. https://thinkchristian.reframemedia.com/should-we-use-a-confession-app

Sawe, Benjamin Elisha. 2017. "Religious Beliefs in Bangladesh." *WorldAtlas*, Apr. 25, worldatlas.com/articles/religious-beliefs-in-bangladesh.html

Sears, Mark. 2018. "AI Bias and the 'People Factor' in AI Development." Forbes, November 13. https://www.forbes.com/sites/marksears1/2018/11/13/ai-bias-and-the-people-factor-in-ai-development/#555088e59134

Sherwood, Harriet. 2017. "Robot Priest Unveiled in Germany to Mark 500 Years Since Reformation." *Guardian,* May 20. https://www.theguardian.com/technology/2017/may/30/robot-priest-blessu-2-germany-reformation-exhibition

Shoemaker, William J. 2012. "The Social Brain Network and Human Moral Behaviour." *Zygon: Journal of Religion and Science* 47: 806-20.

Smith, Megan, Dj Patil and Cecilia Muñoz. *2016. "Big Risks, Big Opportunities: The Intersection of Big Data and Civil Rights."* Obama White House (blog), *May 4*. https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

Stark, Rodney and Willian Sims Bainbridge. 1987. *A Theory of Religion*. New York: David Lang.

Szostak, Rick. 2014. "Classifying for Social Diversity." *Knowledge Organization* 41: 160-70.

Tatlow, Didi Kirsten. 2016. "A Robot Monk Captivates China, Mixing Spirituality with Artificial Intelligence." *New York Times,* April 27. https://www.nytimes.com/2016/04/28/world/asia/china-robot-monk-temple.html

Vidal, Denis. 2007. Anthropomorphism or Sub-Anthropomorphism? An Anthropological Approach to Gods and Robots. *The Journal of the Royal Anthropological Institute* 13: 917-33.

World Economic Forum. Global Future Council on Human Rights 2016-2018. 2018. "How to Prevent Discriminatory Outcomes in Machine Learning." http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

Woodward, James. 2016. "Emotion Versus Cognition in Moral Decision-Making: A Dubious Dichotomy." In *Moral Brains: The Neuroscience of Morality*, ed. S. Matthew Liao. Oxford: Oxford University Press, 87-118. doi:10.1093/acprof:oso/9780199357666.003.0004