Global Faculty Initiative

The Faculty Initiative

seeks to promote the integration

of Christian faith and academic disciplines

by bringing theologians into conversation with scholars

across the spectrum of faculties

in research universities

worldwide.

www.globalfacultyinitiative.net

Disciplinary Brief

VIRTUE ETHICS AND DEVELOPMENT OF AN ETHICAL AI FOR SOCIAL GOOD

Jocelyn S. Downey

Consultant, Department of Electrical and Electronic Engineering, University of Hong Kong

Victor O.K. Li

Chair Professor of Information Engineering, and Cheng Yu-Tung Professor in Sustainable Development, Department of Electrical and Electronic Engineering, University of Hong Kong

Jacqueline C.K. Lam

Associate Professor, Department of Electrical and Electronic Engineering, University of Hong Kong

Jennifer Herdt's Theology Brief on The Virtues reflects on the value of virtue ethics for Christians, the breadth of sources from which virtues may be derived, and the challenges that they raise. Among the virtues discussed by Herdt, the three theological virtues of Faith, Hope, and Love are relevant to the discussions on the ethics of AI. Here, we examine the need for implementing ethical decision-making within Artificial Intelligence (AI) tools and consider the feasibility of applying a virtue ethics model within the AI context and explore how this might be undertaken.

Introduction: the need for ethics in AI

Artificial intelligence (AI) broadly encompasses technologies that are designed to autonomously act in a rational way that, in some sense, mimics humans. A more specific definition has been provided by Russell & Norvig, who confined AI to 'agents that perceive precepts from the environment and take action' [1]. This can involve the abilities to:

- process language naturally (something that ChatGPT has proven successful at),
- perceive or manipulate objects,
- be adaptive and learn from new data (something often referred to as 'machine learning'), and
- find trends in and reach its own conclusions from processed data [2].

A clear need for embedding ethical systems within AI technologies has recently been highlighted by concerns that:

1. Self-driving vehicles face making judgements that have ethical implications [3] [4];

2. ChatGPT has been shown capable of providing solutions to problems that have ethical implications, including offering detailed information on harmful technologies to enquirers;
3. Open-source AI has been used for the purposes of child exploitation [5];
4. The use of AI as a tool for information management has ethical implications [6], with Michael Cuellar focusing on the need to 'discern the kernel of Truth in the midst of all the chaff' [7].

Some of these issues raise further ethical questions regarding the possibility of large-scale social engineering, an issue that has already been raised on existing social media in relation to information management systems [8] [9]. However, the application of an ethical system to any AI must be chosen with care. Across the planet there exists a diversity of philosophies which have produced ethically incongruent systems that diverge even in basic principles such as the value of an individual's life. The existence of this spectrum has implications for the development of ethical AI. The ethics of any system may, ultimately, be dictated by its developers and the context of the system's training. That may embrace the community within which an AI ethical system is trained and those who would exploit its vulnerabilities.

Top-down versus bottom-up in AI ethics

When considering the method of constructing AI ethics, there is a debate over whether, in any given AI, the implementation of an ethical system should take a 'top-down' or 'bottom-up' approach. In 'top-down', ethical principles (such as the Ten Commandments) are imposed on an AI by developers. By contrast, in a 'bottom-up' approach, an AI develops an ethical system by being trained to interpret real-life data, such as the behaviour of vehicle-drivers [10]. Neither system is flawless: a top-down approach is not only subject to the ethical mores of its developmental context, but also runs the risk of being incomplete, and vulnerable to poorly resolving 'grey' areas. On the other hand, a bottom-up approach does not require an AI to be given a set of rules, relying instead on a set of training data to create an ethical system. However, a bottom-up ethical system cannot develop truly *ex nihilo*. It must be programmed with a set of interpretative principles and, since the world is far from perfect, training data for an ethical system may produce poor decision-making. Without due care, it may even produce decisions that contravene legal frameworks, especially because, as Kierkegaard has pointed out, individuals tend to self-justify their own actions [11]. Thus, if an AI were trained using data from a community where there exists an ethical or legal system that no one, in practice, adheres to, the AI would be prone to developing a flawed set of ethical principles or to present opportunities to examine unattainable ethical principles. Furthermore, this system would be vulnerable to developing utilitarian ethics and to exploitation by interest groups who may seek to take advantage of the system by over-loading its interpretative framework with data that biased it in a way that favours the interest group. In practice, therefore a combination of a top-down and bottom-up approach to ethics may produce a more favourable outcome.

Virtue ethics in an AI framework

Approaches that impose rules on the system are sometimes called 'deontological'. An approach where the AI accumulates information on the outcomes of given actions to evaluate whether a given action is beneficial, is called 'consequentialist'. In the search for a framework, it has been argued that both approaches are incapable of keeping pace

with rapid developments in AI across multiple fields [12].

Instead, Terry Bynum [13], Michael Cueller [14] and Jennifer Herdt [15] amongst others, have proposed that a virtue ethics framework would provide a fruitful framework for development. This system of ethics places an emphasis on developing a particular set of character traits rather than making ethical decisions based on a set of rules or the consequences of a particular action. It has recently seen a resurgence in popularity following work by people such as Stanley Hauerwas and Alisdair MacIntyre [16]. Crucially, however, the list of virtues that may be of value for AI systems has not been settled, and the various sets of virtues can have a significantly different focus.

Thus, Aristotle wrote in *Nicomachean Ethics [17]* of the importance of behaving with 'virtue' (ἀρετή/*aretē*), and the Confucian school has produced an ethical framework outlining the importance of acting with 'virtue' (仁/*rén*). Nevertheless, in each case the end is different. Aristotle's principal focus in virtues centres on the importance of 'happiness' (ε*δαιμονία/eudaimonia*), while Confucian texts place a stronger focus on that which produces harmony (和/*hé*) and 'coherence' (理/*lǐ*) within a society. This, perhaps, shares more connections to the Sub-Saharan concept encompassed by the Zulu term '*ubuntu*' (which broadly means, 'shared humanity') than the writings of Aristotle [18] [19], despite sharing the word 'virtue' in English translation.

However, on a practical level, each system contains a number of overlapping ideals and, while each of these cultures has developed geographically separated ideals, each might still exhibit common virtues of a shared humanity. Against a simplistic view of this development, there is strong evidence of cultural interaction between Western Europe, China and trans-Saharan Africa across history, and we cannot exclude the possibility that ideals were passed between cultures. This is seen, for example, in the influence of Hellenism on Buddhist art [20] [21] [22], the presence of Christianity in Tang dynasty China [23], and trans-Saharan trade with Rome during the classical period [24]. More recently, it should be noted that beneath the apparent secularism of Western countries lies a substratum of Judeo-Christian morality. And, due to the historic reach of these countries on the world stage (such as through colonialism, medical missions and the foundation of educational establishments), ethical views that are familiar to Christians can be found in countries where their presence might not be expected.

Historically, the Platonic virtues of Temperance, Courage, Prudence and Justice were assimilated into the Christian world [25]. This was due, in no small part, to the influence of early Christian writers like Ambrose and Augustine [26] (the latter would have been keenly aware of Cicero's writing on the virtues), as well as their inclusion in deuterocanonical texts that show strong Hellenistic influence, such as the Wisdom of Solomon 8:7 *[27]*. The extent of their influence in subsequent centuries can be seen in the depictions of these 'cardinal' virtues within medieval Italian art (see, for example, [28]).

While these virtues had a wide reach in Western society, they are not easily integrated within an AI tool, and they do not naturally map on to the Tanakh and New Testament. The Hellenistic virtue of 'courage' (ἀνδρεία/*andreia*), therefore, possesses a pedigree reaching back to the warrior culture of Homer's men and gods with connotations that do not naturally arise from Biblical texts [29].

Comparing the list of virtues that were used in the classical world with those that can be derived more directly from biblical texts, such as Faith, Hope and Love, Jennifer Herdt, has rightly noted 'there is no exhaustive list of virtues.' She

argues that the virtues are interconnected with one another, and no single virtue can be developed in isolation from the others [30]. Indeed, in Galatians 5:22-23 we can see Paul's thought is in some agreement with this as he provides an interlinked list of virtues ('Love, joy, peace, forbearance, kindness, goodness, faithfulness, gentleness and self-control') as the fruit of the Spirit.

More recently, Cowls and Florid have proposed a different set of virtues for use within an ethical AI framework, focused on the values of autonomy, beneficence, non-maleficence, justice and explicability [31] [32]. These have the advantage of being more easily applied to AI technology and, though they neglect issues of data security and lack a degree of specificity, they provide interesting text virtues with the potential to be relevant to a wide range of applications.

The emergence and derivation of Virtues

However, a question arises: to what extent does nature itself facilitate development of a bottom-up ethical framework when trained on big data taken across diverse languages and cultures? Can virtues emerge from within an AI system, and can these reflect a created order? Many individuals who were instrumental in the development of the modern scientific method had an underlying view that creation is ordered, and that this order can be revealed through investigating creation, and that this order is consistent with the hand of the Judeo-Christian God [33] [34] [35]. In other words, the world is, as Nigel Biggar has noted, 'the creation of a rational Creator' [36]. In another view, with a provenance that reaches at least as far back as the Roman world [37] [38], human life is envisioned as the culmination of competitive and cooperative processes imposed upon random chemical events, from which mind and ethics develop as emergent properties [39] [40] [41].

Whatever processes resulted in human creation, the Judeo-Christian view is that humans were imbued with the 'image of God' (□□□□□□□□ □□□□□ /*tselem elohim*). Though the term 'image of God' may not equate to 'ethical behaviour' it must include the assumption that human behaviour should reflect that expected by God (e.g., Genesis 1.26-28; James 3.9) as well as respect for human life (Genesis 9.6). Observations on the behaviour of the human world may produce various contradictory behaviours concerning what is considered ethical (amplified by settings with legal systems have been framed in ways that have developed in part as a reaction to the one or other worldview, such as Christianity). However, it is possible that behind the maelstrom of data is a distillable core from which would emerge a series of virtues that reflect a consistent ethical order. Indeed, while reported instances of emergence in large language models such as ChatGPT may be no more than 'mirages' [42], machine learning tools are nevertheless capable of deriving the expression of a set of behaviours developed from observance of the natural world. Were a set of universal values proven to be derivable from the study of vast amounts of data, such a system might provide the basis for a universal bottom-up approach for developing ethical AI.

It is, however, highly unlikely that a comprehensive and reliable set of virtues can be derived from this approach. Rather, it will be necessary to impose upon the system values derived from another source. In doing this we must ask not only what criteria we choose from and whether they can be applied across cultures, but also: are the values stable? Will they stand the test of time? As Francis Schaeffer and C. Everett Coop once observed, 'One era is quite certain intellectually and emotionally about what is acceptable. Yet another era decides that these "certainties" are unacceptable and puts together

another set of values into practice. On a humanistic base, people drift from generation to generation, and the morally unthinkable becomes the thinkable as the years move on.' [43]. Beneath many observed changes in values over time lies a single foundational value: autonomy. This is in tension with 'duty' and, whether it is put down to sin or nature, cannot stand as an unrestrained virtue.

In contrast to this, therefore, a set of virtues that emerges from Scripture and could be considered for application to an AI would be Truth and Love. Love equates, not to affirmation or to the promotion of tolerance, but rather to 'other-centredness' – encompassing critical engagement with users with a non-maleficent goal and a duty to protect not only the user and their data but also those whose data is being accessed. These 'virtues' are far from being universally accepted in the world. Hinduism can emphasise the possibility of multiple, potentially contradictory, narratives being equally valid, while within mainstream Islam 'submission' (the meaning of the root *slm*) takes precedence over Love or reasoned Truth. Within the virtues presented here, we can subsume humanity's 'Image of God' within 'Truth' and important ideas such as 'Justice' and 'Mercy' might be seen as an outworking of 'Love.'

Applications of virtue ethics in AI systems for social good

Whether it is possible to distil a set of virtues that have universal application across the spectrum of potential AI technologies is open to debate. Even more of an issue is whether a virtue-ethics system can be implemented in practice. The 'on-the-ground' reality is that states operate on rules (whether by extending 'rights' to people or by limiting rights through a series of 'duties'), and that virtue-ethics-driven AIs may therefore run a high risk of breaching local or international law. Moreover, there is a great danger that any distilled set of virtues, even when implemented in a bottom-up strategy, will be unable to provide the specificity needed for each AI tool, or may even find different virtues stand in tension with each other. The way forward, therefore, may be to develop a hybrid system, whereby virtues are constrained by a series of hard consequences or rules.

If we think briefly about the list of virtues proposed by Cowls and Florid: if an AI is trained to adjust its behaviour according to the virtues of beneficence, non-maleficence, justice, and explicability, in each case a judgement is needed through trained data provided with given rules, as to whether a behaviour is beneficent, just, non-maleficent or explicable. For instance, an AI tool is being trained on open-source image data, to create images according to a user's demands – and the user asks the AI to create a caricature of a named person, (e.g., 'Alex'), that is deliberately seeking to denigrate that person, with the intention that this image can then be posted on social media. Several ethical questions may be posited:

1. How old is Alex? Is there a minimum age for which an image like this can be created? Can the AI accurately judge the age of an individual depiction, and does it matter how photo-real the image is?
2. Is there potential that this will be used in bullying or exploitative material? Were the AI to create such an image it cannot be said to be acting in a non-maleficent or beneficent manner.
3. Is this image in the 'public interest'? It may be that Alex is in the public sphere and this is a legitimate use of his image in satire, such that, while the AI might be breaching its design to be beneficent and non-maleficent, it may be acting justly.

For each of these areas it is most likely that an AI will need to be given information according to local and international law and, while virtue ethics may yet prove to have a great part to play in AI development, we may have to contend with hybrid models. Foundational to any ethical system embedded into an AI must be Truth, such that the ethical model does not simply flutter in the winds of human opinions, or the social mores of a particular generation. Truth can seem elusive: Nietzsche spoke very critically about the reality of Truths [44], and more recently *The New York Times* published an article discussing why young people today do not believe in moral facts [45].

However, while we must not downplay the difficulties of obtaining Truths, we should not abandon the effort. How this will be implemented in practice is the subject of ongoing work within our research group. Below we will showcase how to imbue Truth and Love in the process of building AI for Social Good. First, given the existence of a Truth within a created order [46] which has been created by the epitome of Love (1 John 4.6), our AI system can be designed to mimic such a system. For instance, in designing an automated system for driving, apart from training our model based on real life behavioural data, we can imbue the system with an outworked understanding of the virtue of Love through the Commandment, “Thou shall not kill.” That means, if one’s car is in danger of running into a wall, even though it may be a human tendency to swerve and hit pedestrians or other drivers in order to save one’s own life, it might be necessary from a Christian perspective that such an outcome generated by an AI decision-making system should be avoided or taken as a last resort.

Truth is not derived through a consensus or democratic decision (though the recognition of that Truth might be). For good or ill, it will ultimately be incumbent upon experts in AI to take a lead on ethical development of AI, and Christians within the field of AI must share with both their colleagues and the public because such a Christian-based ethics approach can be desirable for the prosperity and sustainability of the society over the longer term. In other words, AI should carefully avoid being misguided by the inherent order exhibited by data, which may likely reflect the worldviews and behaviours of the sinful as well as Christian virtues. In this regard, it will be important to consider the role of biblically driven ethics that might guide the future development of advanced AI models, such as Large Language Models (LLMs), or Generative AIs (GAIs). Truth must not be compromised no matter how sophisticated these models will become.

For example, while acknowledging how powerful LLMs can perform linguistic or other different types of tasks such as coding, we must share with the community the risks that LLMs, such as ChatGPT or Bard, can potentially bring to our societies and individuals such as hallucination or stealing of personal information, and propose sensible solutions. Lastly, humans, as possessors of God’s image, are the ultimate managers of the earth, rather than vice versa. An AI-system governed by Truth and Love is intended to assist human beings in exercising their roles even more fully as stewards of the earth (Genesis 1.26) rather than becoming being ruled by it. Any AI system developed to control humans works against this principle and should be prohibited at all costs.

End Notes

[1] Russell, S. J., and Norvig, P. (2010) Artificial Intelligence: a modern approach, 3rd edition ed., Prentice Hall, Upper Saddle River, NJ

[2] Ibid.

[3] Maxmen, A. (2018) Self-driving car dilemmas reveal that moral choices are not universal. *Nature* **562**, 469-470

[4] Rhim, J., Lee, J. H., Chen, M., and Lim, A. (2021) A Deeper Look at Autonomous Vehicle Ethics: An Integrative Ethical Decision-Making Framework to Explain Moral Pluralism. *Front Robot AI* **8**, 632394

[5] Anderljung, M., and Hazell, J. (2023) Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? *ArXiv* **abs/2303.09377**

[6] Eliot, L. (2022) AI ethics and AI-induced psychological inoculation to help humans with disinformation. in *Forbes*

[7] Cuellar, M. (2023) A virtue ethical approach to the use of artificial intelligence. *Data and Information Management*, 100037

[8] Gisondi, M. A., Chambers, D., La, T. M., Ryan, A., Shankar, A., Xue, A., and Barber, R. A. (2022) A Stanford Conference on Social Media, Ethics, and COVID-19 Misinformation (INFODEMIC): Qualitative Thematic Analysis. *J Med Internet Res* **24**, e35707

[9] Flores, L., and Young, S. D. (2022) Ethical Considerations in the Application of Artificial Intelligence to Monitor Social Media for COVID-19 Data. *Minds and Machines* **32**, 759-768-

[10] Etzioni, A., and Etzioni, O. (2017) Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* **21**, 403-418

[11] Kierkegaard, S. (2008) *A Sickness Unto Death*, Penguin

[12] Cuellar, M. (2023) A virtue ethical approach to the use of artificial intelligence. *Data and Information Management*, 100037

[13] Bynum, T. (2008) Norbert Wiener and the rise of information ethics. in *Information Technology and Moral Philosophy* (Hoven, J., and Weckert, J. eds.), Cambridge University Press, Cambridge. pp 8-25

[14] Cuellar, M. (2023) A virtue ethical approach to the use of artificial intelligence. *Data and Information Management*, 100037

[15] Herdt, J. (2023) The Virtues. Global Faculty Initiative, https://globalfacultyinitiative.net/content_item/209

[16] Cloutier, D., and Mattison, W. C. (2014) The Resurgence of Virtue in Recent Moral Theology. *J Moral Theol* **3**, 228-259

[17] Aristotle, and Crisp, R. (2000) *Nicomachean ethics*, Cambridge University Press, Cambridge, U.K.

[18] Bell, D. A., and Metz, T. (2011) Confucianism and ubuntu: Reflections on a dialogue between chinese and african traditions. *Journal of Chinese Philosophy* **38**, 78-95

[19] Metz, T. (2020) Ubuntu: The Good Life. in *Encyclopedia of Quality of Life and Well-Being Research* (Maggino, F. ed.), Springer International Publishing, Cham. pp 1-5

[20] Kubica, O. (2023) *Greco-Buddhist relations in the Hellenistic Far East. Sources and Contexts*,

[21] Rienjang, W., and Stewart, P. (eds). (2019) *The global connections of Gandharan art : proceedings of the Third International Workshop of the Gandhara Connections Project, University of Oxford, 18th-19th March, 2019*

[22] Aldrovandi, C., and Hirata, E. (2005) Buddhism, Pax Kushana and Greco-Roman motifs: pattern and purpose in

Gandharan iconography. *Antiquity* **79**, 306-315

[23] Liščák, V. (2009) Early Chinese Christianity in the Tang Empire: On the Crossroads of Two Cultures. pp 103-125

[24] Schörle, K. (2012) Saharan Trade in Classical Antiquity. in *Saharan frontiers : space and mobility in Northwest Africa* (McDougall, J., and Scheele, J. eds.), Indiana University Press, Bloomington. pp 58-72

[25] Herdt, J. (2023) The Virtues. Global Faculty Initiative, https://globalfacultyinitiative.net/content_item/209

[26] Kramer, C. (2015) *Augustine on Christian and Pagan Virtues (2011)*,

[27] Barrier, J. W. (2023) Middle Platonism in the Wisdom of Solomon: A comparison of Wisdom to Plutarch of Chaeroneia. *Journal for the Study of the Pseudepigrapha* **32**, 244-269

[28] Watanabe, Y. (2022) Filippo Lippi's Frescoes at Spoleto, Cardinal Eroli, and the Immaculate Conception. *Mitteilungen des Kunsthistorischen Institutes in Florenz* **64**, 221-229

[29] Zavaliy, A. (2017) How Homeric is the Aristotelian Conception of Courage?: Homeric and Aristotelian Conceptions of Courage. *The Southern Journal of Philosophy* **55**, 350-377

[30] Herdt, J. (2023) The Virtues. Global Faculty Initiative, https://globalfacultyinitiative.net/content_item/209

[31] Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019) The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, Honolulu, HI, USA

[32] Cowls, J., and Floridi, L. (2018) Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. SSRN

[33] Goodman, D. C. (1973) *Science and religious belief, 1600-1900. (A selection of primary sources.) Edited by D. C. Goodman*, J. Wright, Bristol

[34] Jones, R. H. *For the glory of God : the role of Christianity in the rise and development of modern science*, Lanham : University Press of America, c2011-

[35] Hannam, J. (2009) *God's Philosophers: How the Medieval World Laid the Foundations of Modern Science*, Icon Books Ltd, London

[36] Biggar, N. (2023) Order. Global Faculty Initiative, https://globalfacultyinitiative.net/content_item/218

[37] Tertullian, and Felix, O. (1953) *Apology, De Spectaculis*, Harvard University Press

[38] Cicero, M. T., and Ross, J. M. (1972) *The nature of the gods*, Penguin, Harmondsworth,

[39] Milinski, M. (2011) Biology: A revolution in evolution. *Nature* **471**, 294-295

[40] Burkart, J. M., Brügger, R. K., and van Schaik, C. P. (2018) Evolutionary Origins of Morality: Insights From Non-human Primates. *Frontiers in Sociology* **3**

[41] Polanyi, M., and Sen, A. (2009) *The Tacit Dimension*, University of Chicago Press, Chicago, Ill.

[42] Schaeffer, R., Miranda, B., and Koyejo, S. (2023) Are Emergent Abilities of Large Language Models a Mirage? *ArXiv* **abs/230415004v2**

[43] Schaeffer, F. A., and Koop, C. E. (1979) *Whatever Happened to the Human Race?*, Fleming H. Revell Company

[44] Bittner, R. (2021) Nietzsche's Concept of Truth. in *Nietzsche as German Philosopher* (Höffe, O. ed.), Cambridge University Press, Cambridge. pp 231-254

[45] McBrayer, J. P. (2015) Why Our Children Don't Think There Are Moral Facts. in *The New York Times*, March 2nd 2015

Ed., New York

[46] Biggar, N. (2023) Order. Global Faculty Initiative, https://globalfacultyinitiative.net/content_item/218

For more information

www.globalfacultyinitiative.net