Agathon: A Journal of Ethics and Value in the Modern World, Vol 10/2023

Chapter 15
Confucian Role Ethics and Artificial Intelligence

Lawrence A Whitney

'To be able under all circumstances to practice five things constitutes perfect virtue; these five things are gravity, generosity of soul, sincerity, earnestness, and kindness.'
—Confucius[1]

Abstract: Among virtue ethics frameworks, I argue that Confucian approaches, sometimes called role ethics, have an easier time incorporating artificial intelligence (AI) into an analysis of virtue than Western theistic approaches. This stems in part from differences in the ethical theories themselves, namely the availability of distinctive virtues for particular roles in Confucian approaches, but also because of the philosophical/theological anthropologies they presuppose regarding the value of changing our endowed human nature. I describe the emergence of role ethics in the context of ancient Chinese philosophy, and compare key concepts such as 仁 (*ren*; humaneness), 義 (*yi*; appropriateness), and 孝 (*xiao*; filialty) with Aristotle and his Christian, Jewish, and Muslim inheritors. This forms the basis for an applied analysis of three types of AI: self-driving cars, large language models, and neural implants.

Key Terms: Confucianism; Roles; Virtue; Self-driving Cars; Neural Implants

The authors of *The AI Revolution in Medicine* note that the artificial intelligence (AI) known as GPT-4 'is at once both smarter and

1. Khurana, Simran, '47 Confucius Quotes That Still Ring True Today', ThoughtCo, Apr. 5, 2023, thoughtco.com/best-confucius-quotes-2833291 (accessed 2/22/2024).

dumber than any person you've ever met'.[2] In addition to provoking the longstanding and ongoing question of whether AIs are in fact intelligent, this observation relies on a fundamental and irreducible difference between AIs and human persons.

Regardless of precisely what this difference is taken to be, or even its degree, its recognition raises ethical questions as to whether the measures of human behavior elaborated in ethical theories are properly applicable to AIs. Indeed, until recently, ethicists have decisively claimed that only humans are capable of moral reason and therefore can be held accountable for behaving ethically, though Mark Rowlands' *Can Animals be Moral* has vigorously shaken that foundation.[3] Having relied on this circumscription of their domain to the realm of human behavior, ethical theorists have largely been able to construe their theories in universal terms, applying to everyone, everywhere, always. With the capacity for moral reasoning now on the table for animals, and perhaps even more so for AIs, it becomes necessary to consider a form of ethical pluralism that incorporates not only the potential conflict among ethical norms and values but also differences among the norms and values applicable to different morally reasoning entities. Here I argue that a Confucian formulation of virtue ethics, or role ethics,[4] serves as a useful entre into developing such an ethical pluralism that might be profitably applied to nonhuman agents, including AIs.

In making the case for the Confucian approach to virtue ethics, I begin by elaborating the vocabulary and frameworks of three Warring States period (453-221 BCE) thinkers: 孔子 (Kongzi or 'Confucius'; ca. 551-479 BCE), 孟子 (Mengzi or 'Mencius'; ca. 372–289 BCE), and 荀子 (Xunzi; ca 310–220 BCE). Of particular interest here are the differences among which virtues apply and how they are to be lived out with respect to different personages inhabiting the orbit of a given moral agent.

2. Quoted in Hilary Brueck, 'ChatGPT Can Save Lives in the ER, but It Needs Supervision: 'It Is at Once Both Smarter and Dumber than Any Person You've Ever Met'", in *Insider* (blog), April 7, 2023, https://www.insider.com/chat-gpt-successor-gpt-4-can-help-doctors-save-lives-2023-4; Peter Lee, Carey Goldberg, and Isaac Kohane, *The AI Revolution in Medicine* (New York, NY: Pearson Education, 2023).
3. Mark Rowlands, *Can Animals Be Moral?* (Oxford: Oxford University Press, 2012).
4. Roger T Ames, *Confucian Role Ethics: A Vocabulary* (Honolulu, HI: University of Hawai'i Press, 2011).

This approach contrasts with the dominant strain of Western virtue ethics: eudaimonist virtue ethics, which emerged from Aristotle and has been appropriated in Christian (for example, Aquinas, 1225–1274 CE), Islamic (for example, Miskawayh, 932–1030 CE), and Jewish (for example, Maimonides, 1138–1204 CE) traditions. Comparison of these two virtue ethics lineages forms the basis for the claim that an ethical pluralism underwritten by the Confucian approach is better able to accommodate nonhuman agents. This is so first because the Confucian conception of distinct roles includes universal virtues applicable to all, but also allows for particular virtues associated with discrete roles in the social matrix. Second, whereas the Aristotelian lineage as adopted in Abrahamic traditions conceive virtuous agents having a fixed, divinely endowed nature, Confucian virtue ethics are rooted in a philosophical anthropology in which changing, i.e. improving, the baseline nature of a moral agent is the whole point of moral self-cultivation toward virtue. Nonhuman agents, such as AI systems, may thus inhabit distinct roles with discrete virtues associated with them, and may thereby participate in the process of humans becoming more virtuous, along with the wider process of achieving a virtuous society.

Confucian Ethics

Confucian ethics emerged very much as the core of the Confucian intellectual project in the context of rampant socio-political fracturing and disintegration at the end of the Spring and Autumn (771–453 BCE) and throughout the Warring States periods in Chinese history.[5] For Confucius and the scholar-officials who took inspiration from him, ethics was key to a political philosophy that sought to restore social stability. Confucius began by reinventing the conception of the 君子 (*junzi*; noble person) on moral, rather than class, terms: 'Originally, the meaning of the term *junzi* was 'son of a lord'", but for Confucius, 'the *junzi* is less the noble man whose nobility derives from inherited *social* nobility than the noble person whose nobility

5. Yuri Pines, *Foundations of Confucian Thought: Intellectual Life in the Chunqiu Period, 722-453 B.C.E.* (Honolulu, HI: University of Hawaii Press, 2002); Yuri Pines, *Envisioning Eternal Empire: Chinese Political Thought of the Warring States Period* (Honolulu, HI: University of Hawaii Press, 2009).

derives from personal commitment and a developed *moral* power'.[6] Just as the English word 'noble' now denotes a person who exhibits a particular virtue or set of virtues, the concept of the junzi for Confucians might be translated as 'virtuous person'.

Likewise, Confucius set about renovating the conception of the virtue that such noble persons embody, which is 仁 (*ren*; humaneness). *Ren* 仁 was the stative verb form of *Person Ren* 人, "which the aristocratic clans of Zhou used to distinguish themselves from the common people . . . The noble, civilized, fully human, pride themselves on their manners and conventions, but above all on the virtues which give these meaning and which distinguish themselves from the boors and savages who do not know how to behave."[7] Confucius sought to meritocratise the notion of ren so that it would extend beyond the noble class to include his own ministerial class of 士 (*shi*; scholar-apprentices).

For Confucius, *ren* is the conception of virtue in general, whereas the junzi is the person who embodies it. This is to say that '*ren* is an at-large virtue; only an individual commitment to it makes it a personal dao, or principle, for that person'.[8] This concept of 道 (*dao*; way or principle) for Confucius is of an individual, personal, moral principle, which is almost precisely inverse from the conception subsequent Confucians would come to have of *dao* as a cosmological, universal, metaphysical principle, largely developed in dialogue and debate with their frequent competitors, the Daoists. 'The word dao 道, originally meaning "way" or "road", is used everywhere by the philosophers to mean the way to do something, or the (right moral) "Way", or (later) the "Way" of all nature.'[9] For later thinkers, *dao* is transformed from an internal principle to an external norm, but 'for

6. William Theodore de Bary and Irene Bloom, eds., *Sources of Chinese Tradition: Volume 1: From Earliest Times to 1600* (New York, NY: Columbia University Press, 1999), 42.
7. Angus Charles Graham, *Disputers of the Tao: Philosophical Argument in Ancient China* (La Salle, IL: Open Court, 1989), 19.
8. Confucius, 論語辨 *The Original Analects: Sayings of Confucius and His Successors*, translated by E Bruce Brooks and A Takeo Brooks (New York, NY: Columbia University Press, 1998), 14.
9. See *The Cambridge History of Ancient China: From the Origins of Civilization to 221 BC,* edited by Michael Loewe and Edward L Shaughnessy (Cambridge, UK: Cambridge University Press, 1999), 750–51.

Confucius, *dao* is primarily *rendao* 人道, that is "a way of becoming consummately and authoritatively human'",[10] for nobles or ministers alike. Confucius thus considers *ren* to be the universal principle of virtue, whereas *dao* is the set of particular virtues that the *junzi* puts into practice in daily life.

By the later part of the Warring States period, inhabitants of the Confucian lineage would make significant changes to this ethical vision. Mencius, for example, derives *ren* as one of four cardinal virtues alongside 義 (*yi*; appropriateness), 禮 (*li*; ritual propriety), and 智 (*zhi*; wisdom).[11] *Ren* is thus no longer an at-large, general virtue, but rather a specific virtue that derives from instinctual human feelings rooted in an essential human nature that is fundamentally good. Xunzi, who takes human nature to be more ominous, retains the generality of *ren* as a virtue, but charts the path to achieving it not through cultivation of subsidiary virtues but by transforming that nature through education and ritual (Xunzi 1988, 2.11, volume 1:157). 禮 (*li*; ritual) here is not a virtue but a regime of practices and behavioral patterns that generate bearings and dispositions, which in time make their practitioners (*cheng*; sincere) in the humaneness they reflect.[12]

This focus on rule-governed transformational practices have led some scholars to conclude that Confucian ethics are not a type of virtue ethics at all, but rather a deontological ethics more properly in conversation with the Kantian project, for example Mou Zongsan.[13] Others have questioned whether Confucian ethics fit into any of the three primary Western paradigms, (virtue, deontology, consequentialism), while acknowledging affinities with less dominant

10. Confucius, *The Analects of Confucius: A Philosophical Translation*, translated by Roger T Ames and Henry Rosemont, Jr, 1st edition (New York, NY: Ballantine, 1999), 46.
11. Mengzi, *Mengzi: With Selections from Traditional Commentaries*, translated by Bryan W Van Norden (Indianapolis, IN: Hackett, 2008), 2A6.7, 149.
12. Xunzi, *Xunzi: A Translation and Study of the Complete Works*, translated by John Knoblock (Stanford, CA: Stanford University Press, 1988), 3.9a, volume 1:177; Yanming An, *The Idea of Cheng (Sincerity/Reality) in the History of Chinese Philosophy* (New York, NY: Global Scholarly, 2005), 48.
13. Kam-por Yu, Julia Tao, and Philip J Ivanhoe, *Taking Confucian Ethics Seriously: Contemporary Theories and Applications* (State University of New York Press, 2010), 27–52, 73–98.

strains of ethical theorizing such as social ethics.[14] Much of the challenge here has to do with core questions in comparative method, especially the issue of which comparator sets the controlling discourse. Rather than attempting an *a priori* determination of the possibility of comparison, this pitfall is often best avoided by undertaking the comparison and then adjudicating its fruitfulness.

Perhaps the most distinctive feature of Confucian ethics, especially as form of virtue ethics, is the way in which virtues are variously expressed through the roles a moral agent inhabits in relation to others in their orbit. The concept of role is not explicitly elaborated in Confucian texts but is rather a contemporary interpretation of how Confucian ethics conceives of virtuous behavior being worked out in diverse social circumstances. Historically speaking, the concept of role is best understood as emerging from a process of abstracting, transposing, and systematizing the concept of 孝 (*xiao*; filialty), which originally had to do with "honor and obedience to one's parents."[15] What this honor and obedience entailed varied based on the gender and birth order of the child in question, and thus the relationships between different children and their parents were characterized by distinct duties, responsibilities, and behavioral patterns. Even in the Warring States period, filiality began to be transposed from the domain of family relations to a much wider set of relationships:[16] 'It is clear that filial devotion can be translated into political loyalty, professional dedication, personal trustworthiness, and even military courage.'[17]

Filiality, then, becomes something like behaving in ways that are distinctly appropriate (義 *yi*; appropriateness, rightness) with respect to each person in your social orbit, which in turn is the achievement of *ren*, and thus the becoming of a *junzi*. To be a moral agent in this schema is to inhabit the various roles determined by the network of social relations in which the agent is situated by behaving in ways that fulfill the norms that govern each relationship.

14. AT Nuyen, 'Confucian Ethics as Role-Based Ethics', in *International Philosophical Quarterly* 47, no 3 (2007): 315–28, https://doi.org/10.5840/ipq200747324.
15. Loewe and Shaughnessy, *The Cambridge History of Ancient China*, 479.
16. Keith N. Knapp, 'The Ru Reinterpretation of *Xiao*', in *Early China,* 20 (1995): 195–222, https://doi.org/10.1017/S036250280000448X.
17. *Dao Companion to Classical Confucian Philosophy,* edited by Vincent Shen (Dordrecht: Springer, 2014), 110.

Eudaimonist Virtue Ethics in Comparison

The dominant Western versions of virtue ethics today derive largely from the virtue theory developed by Aristotle (384-322 BCE) in the *Nichomachean Ethics*.[18] For Aristotle, the ultimate goal of life is *eudaimonia*, which means something like 'living well', which is to say that the ultimate goal of life is to live a good life, and so he sets out to explain what that is and how to do it. Virtues, then, are habits cultivated in childhood through social learning and tempered by practical wisdom (*phronesis*) as the moral agent matures. Each ethical virtue is intermediate, or a mean, between an excess and a deficiency of character ascribable to appropriate action, e.g. courage is the mean between the excess of rashness and the deficiency of cowardice.[19] To live well is to put the ethical virtues into practice by applying them in concert with practical wisdom to the situations of daily life.

This framework for virtue ethics has been widely influential across the theistic Abrahamic traditions of Christianity, Judaism, and Islam. Thomas Aquinas (ca 1225–1274 CE), whose *Summa Theologica* remains a touchstone of Christian theological education, largely adopted Aristotelian virtue ethics, though also adapted the approach to make it consistent with his Christian theism. One such adaptation is his development of a theory of moral law, which includes not only divine law revealed to humanity, for eexample in biblical texts, but also natural law inherent in humans having been endowed with reason in divine creation.[20] The appropriation of Aristotelian virtue ethics into Judaism in the figure of Moses ben Maimon (Maimonides; 1138–1204 CE), by contrast, does not so rely on innate natural law to achieve a universalisable ethic, but rather underwrites virtue with revelation interpreted in the light of reason, which is then transmitted universally outward by its practice among

18. Aristotle, *Aristotle: Nicomachean Ethics*, edited by Roger Crisp (Cambridge, UK: Cambridge University Press, 2014); Richard Kraut, 'Aristotle's Ethics', in *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta and Uri Nodelman, Fall 2022 (Metaphysics Research Lab, Stanford University, 2022), https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/.
19. Aristotle, *Aristotle*, 25.
20. Robert Pasnau, 'Thomas Aquinas', in *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta and Uri Nodelman, Spring 2023 (Metaphysics Research Lab, Stanford University, 2023), https://plato.stanford.edu/archives/spr2023/entries/aquinas/.

Jews. Maimonides also disagreed with Aristotle about habit ultimately resulting in a fixed character of goodness, instead requiring ongoing critical self-consciousness to right wrongs and return to virtue.[21] Ahmad ibn Muhammad ibn Miskawayh (ca 940–1030 CE) likewise relied on revelation, though that revealed in the Qur'an, of course. Ibn Miskawayh also developed a fundamentally social conception of virtue as formed in the public sphere, relying on friendships in community,[22] which emphasis has reemerged in the contemporary renaissance of virtue ethics inaugurated by Alasdair MacIntyre.[23]

Each of these thinkers, from three distinct theistic traditions, works to articulate the role and configuration each of revelation and reason in generating the disposition toward virtue characteristic of Aristotelian virtue ethics. Not being beholden to the particularity of revelation, Aristotle was able to straightforwardly rely on reason as a common characteristic of humanity, which commonality makes his ethic universally applicable. The reason Aquinas, Maimonides, and Ibn Miskawayh need to get the relationship between revelation and reason right is precisely to safeguard the universality of virtue ethics as a moral philosophy emerging from under the penumbra of their theistic worldview. The key point here is that the virtues articulated by Western virtue ethicists located in the Aristotelian lineage are intended to apply universally, one and the same across all instances for each and every moral agent. It may be that how the virtues are manifested in behavior vary according to the situation, but every moral agent is expected to acquire and maintain each and all of the same set of virtues and behave in any and all situations according to those means between their respective excesses and deficiencies.

21. Jonathan Jacobs, 'Aristotle and Maimonides on Virtue and Natural Law', in *Hebraic Political Studies* 2, no 1 (Winter 2007): 46–77; Kenneth Seeskin, 'Maimonides', in *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Spring 2021 (Metaphysics Research Lab, Stanford University, 2021), https://plato.stanford.edu/ENTRIES/maimonides/.
22. Elizabeth M Bucar, 'Islam and the Cultivation of Character: Ibn Miskawayh's Synthesis and the Case of the Veil', in *Cultivating Virtue: Perspectives from Philosophy, Theology, and Psychology*, edited by Nancy E Snow (New York, NY: Oxford University Press, 2014), 0, https://doi.org/10.1093/acprof:oso/9780199967421.003.0009.
23. Alasdair C MacIntyre, *After Virtue: A Study in Moral Theory* (University of Notre Dame Press, 1981).

Confucians also have universal virtues that are understood to apply to all moral agents, though in the formative years of the tradition much of the work undertaken by Confucian thinkers had to do with making them universal rather than parochial to the elite social class. As already noted, Confucius extended the concept of humaneness as the telos of virtue for moral agents at least to also include his own class of minor aristocrats as moral agents. Filiality was universalized in another way, extending the scope not of the agent but of the direct object of its purview, from parents to a much broader range of social relations, and eventually encompassing all. Mencius clearly understood his four cardinal virtues to be universal, deriving as they do from a universal human nature of goodness.[24]

At the same time, central to Confucian thought is careful analysis of the norms that govern the behaviors of agents interacting in various social roles. This attention to roles arises in part from the universalizing process applied to the virtue of filiality. Whereas filiality originally applied only to relationships between children and their parents, the process of abstracting and applying the principles of loyalty, deference, and respect to people in other social relations meant recognizing the ways in which the roles those others inhabited were similar to and different from the role of parent. For example, the ruler is in some respects similar to a parent in terms of their responsibility for meeting the needs of the populace but is also very different in that there is no direct care of citizens, and thus little of the intimacy that is so crucial in relationships between parents and children. The loyalty, deference, and respect of a citizen toward their ruler thus looks in some ways similar but in many important ways rather different than the loyalty, deference, and respect a child should display toward their parents. Nevertheless, those different behaviors are still embodiments of the common virtue of filiality.

Given that many of the social roles with attendant norms that Confucians analyze, and view as characteristic of a humane society, themselves predate the movement toward universalization of virtue, it is little wonder that tension emerges from the beginning between those norms and the demands of universalized virtue. By contrast, this tension has only rather recently begun to be identified and

24. RAH King, 'Universality and Argument in Mencius IIA6', in *Proceedings of the Aristotelian Society* 111 (2011): 275–93, https://www.jstor.org/stable/41331551.

explored in Western virtue ethics.[25] John Ramsey helpfully identifies this tension between virtue and norm as the 'role dilemma', which is the conflict between the demands of a social role and the demands of virtue. He distinguishes between responses as either externalist, turning to virtues that exist beyond the scope of the role in question, or internalist, resolving the dilemma by adopting the norms of the role. Ramsey further notes that the former response collapses into a virtue ethics much along the lines of the Aristotelian variety, whereas the latter 'implies a form of cultural relativism and allows for repressive and problematic social institutions'.[26]

Notably, Ramsey understands the notion of role, or at least its norms, to be closely aligned with the concept of 禮 (*li*; ritual). The dilemma he identifies thus puts ritual in tension with virtue in a way that he claims is inadequately addressed among the early Confucians. Since all of the early Confucians extensively explicate their understandings of ritual, humaneness, and their interrelations, it may instead be that the framing of the dilemma in terms of internalism and externalism is the source of difficulty in reconciling the dilemma, and seeing how it is reconciled by these thinkers. Moreover, such a tension is strange in light of Aristotelian virtue ethics because ritual is very similar to the Aristotelian conception of habit, which like ritual is the means of achieving and cultivating virtue.

Critical for reconciling ritual and humaneness for Confucians is the concept of 義 (*yi*; appropriateness), which Mencius identified as another of his cardinal virtues. Rather than identifying them all as virtues, Sor-hoon Tan calls these the 'three key ethical ideas of authoritative conduct (*ren* 仁), appropriateness (*yi* 義), and ritual practice (*li* 禮)'. Adopting a perspective on appropriateness from David Hall and Roger Ames, and grounding it in the texts of Confucius, Mencius, and Xunzi, Tan says that appropriateness 'has to do with the personal investment of meaning in action, based on the interaction between a person's individuality and her environment in specific situations'.[27] This is not dissimilar to what John Knoblock

25. Sean Cordell, 'Virtuous Persons and Social Roles', in *Journal of Social Philosophy* 42, no 3 (2011): 254–72, https://doi.org/10.1111/j.1467-9833.2011.01535.x.
26. John Ramsey, 'The Role Dilemma in Early Confucianism', in *Frontiers of Philosophy in China* 8, no 3 (2013): 377–78, https://www.jstor.org/stable/23597454.
27. Sor-hoon Tan, *Confucian Democracy: A Deweyan Reconstruction* (Albany, NY: State University of New York Press, 2012), 83.

describes in saying that 'Yi expresses the "rightness" of a course of conduct that is proper, fitting, decent, suitable, appropriate in the circumstances in which it was done'. He further describes how the conception of appropriateness emerged from the universalisation process of filiality: 'Yi thus designated the appropriateness, the fitness, and the suitability of the service the minister gave his lord and the son his father, the respect the humble gave the noble, the assistance friends gave each other, and the differences in treatment between near and far relatives.'[28] And yet, citing *Mencius* 6A5,[29] Knoblock notes that appropriateness 'becomes more than mere congruity since it reflects an inner sense for what is right',[30] which is to say virtue. Appropriateness is thus the process, in the moment, of generating harmony amidst the sometimes competing demands of virtue and the ritual norms governing the roles inhabited by those involved in the situation at hand. Appropriateness overcomes the role dilemma Ramsey posits in the act of generating harmony from the tension rather than from an *a priori* determination across instances of the correct balance between humaneness and ritual principles.

Confucian appropriateness is thus akin to Aristotelian practical wisdom (*phronesis*). For Aristotle, 'practical wisdom, as he conceives it, cannot be acquired solely by learning general rules. We must also acquire, through practice, those deliberative, emotional, and social skills that enable us to put our general understanding of well-being into practice in ways that are suitable to each occasion.'[31] Practical wisdom is thus about bringing virtues to life in the concrete situations of daily life. Appropriateness is likewise situational, contextual, and skill-based. What appropriateness also brings to the table, that practical wisdom does not, however, is the set of ritual norms that govern behavior between inhabitants of various roles. These ritual norms are themselves based on prior instantiations of humaneness in encounters between agents, and so there is a dialectical relationship between the ideal of virtue and the concreteness of ritual and role.

Having such prior concrete examples of virtue in action to rely on gives the Confucians a leg up on the Aristotelians in moral decision making. The Confucians can rely on what Daniel Kahneman calls

28. Xunzi, *Xunzi: Translation and Study*, I.95.
29. Mengzi, *Mengzi*, 147–48.
30. Xunzi, *Xunzi: Translation and Study*, I.95.
31. Kraut, 'Aristotle's Ethics'.

system one, or 'fast', thinking, which relies on heuristic, habit, and past patterns to make decisions quickly, rather than having to rely on system two, which is slow, deliberate, and logical.[32] Whereas Aristotle locates habit at the beginning of moral formation, as the source of virtue development, the Confucians keep ritual and habit in the mix all the way through, recognizing that humans have to operate with a bounded rationality much of the time, under conditions of limited knowledge, resources, and time.[33] Appropriateness has to do not only with the rational process of applying virtue to a situation, as for Aristotelan practical wisdom, but also the instinctual recognition of when, where, and how a ritual pattern fits around the participants in an encounter.

One good reason for considering Confucian role ethics to be a form of virtue ethics in relatively close proximity to that of the Aristotelian eudaimonst variety, rather than a distinct ethical type or more closely approximating deontology or consequentialism, is that the norms that govern social roles in the theory may best be interpreted as particular rather than universal virtues. They are certainly, and explicitly, expressions of the broader universal virtues elaborated above, but these norms are also more like virtues than they are like rules as they become sincere expressions of feeling formed through the practice of the rituals that mediate each role, which is very similar to the formation program of habits as conceived by Aristotle. That said, Confucian thinkers do not view that formation program as leading to the learning of general, universal virtues, but rather of the particular virtues that govern the role in question. For example, the virtue of deference is a general virtue that applies to more junior members of families and the state with respect to their respective superiors alike, but that is not to say that learning the appropriate virtue of deference with respect to a parent means that one has also learned a general principle that can be applied to relating to ministers and nobles. Those deferential relationships must be separately learned according to their ritual norms, and only then can the moral agent recognise and appreciate the commonalities between them that is the general virtue of deference. Learning appropriate deference across a

32. Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
33. Gerd Gigerenzer and Reinhard Selten, *Bounded Rationality: The Adaptive Toolbox* (Chicago: MIT Press, 2002).

variety of roles thus hones the meaning of the virtue of deference for the moral agent such that they may more appropriately implement it in each role.

Particular Roles and Virtues for AI

With these comparative considerations in mind, it is now possible to more precisely analyze how the Confucian version of virtue ethics would approach artificial intelligences. Key to note is that at least two forms of particularity accompany the Confucian conception of roles such that the Confucian approach to virtue ethics is best understood as a role ethic.

The first form of particularity is with respect to the role itself such that universal virtues are expressed differently in behavior depending on the role of the moral agent and the roles of those toward whom their behavior is directed. For a child to be respectful toward a parent requires different behavior than for a soldier to be respectful toward a general. As such behavioral patterns become codified and proscribed apart from direct recourse to the virtue they are meant to express, situations may arise in which enacting the behaviors would in fact conflict with the demands of virtue, which is the role dilemma identified by Ramsey as shared between Aristotelians and Confucians.

The second form of particularity is with respect to virtues that are particular to one or a discrete set of roles but are therefore not universal. Courage, for example, is not a universal virtue in Confucianism as it is not expected for many roles, particularly roles inhabited by women, who are associated with 陰 (*yin*; passive or negative). A virtue dilemma may emerge here, where the behaviors that express a particular virtue associated with the role a moral agent inhabits may conflict with the universal virtues applicable to all. For example, sons have particular filial responsibilities for parents, which become especially important as the parents age, but in the case where there is no son, which is increasingly common as a result of the One Child Policy in China, daughters are faced with a conflict between the passivity virtue particular to female roles and the universal virtues of filiality and humaneness.

This second form of particularity is unique to the Confucian approach as it does not have a direct analogue in Aristotelian virtue

ethics. This is because the Aristotelian approach conceives of each virtue applying universally to all moral agents and at least in principle applicable in all situations, whereas the Confucian approach need not conceive all universal virtues as necessarily applicable to every role. While respect is a virtue that applies to both superiors and inferiors in social relations, albeit realized for each according to the specific virtues of their roll, the virtue of deference applies only to inferiors with respect to their interactions with superiors.

This deference differential is demonstrated by the need, from quite early in the tradition, to create a mechanism for dealing with immoral and despotic superiors, especially rulers. In the 孝經 (*Xiaojing*; *Classic of Fillial Piety*), a student asks Confucius whether children must obey every command of their father, as deference and attendant virtues would seem to imply. In response, Confucius details how in the past, rulers at various levels would have officials whose duty was to 諫 (*jian*; remonstrate) with them when they made bad decisions so that they would keep their states on the 道 (*dao*; way). He then draws the analogy with a father behaving immorally such that a child is morally obligated to remonstrate with them, and concluding by demanding "How could simply obeying the commands of one's father be deemed filial?"[34] The mechanism of remonstration is the exception to correct for the risk of corruption and immorality raised by the structural difference in application of the virtue of deference only from inferiors to superiors.

These forms of particularity give Confucians a great deal more flexibility when applying virtue ethics to artificial intelligences (AIs), yet also result in a broader moral topography that requires charting. At the start, like Aristotelians, Confucians need to consider how universal virtues applicable to all moral agents apply to AIs, which are presumed to be moral agents because if they are not then virtue ethics, and arguably all normative ethical paradigms, would not apply. Confucians then have a number of other trajectories of analysis to undertake, which must begin with a conceptualization of the role or roles that AIs inhabit in the social sphere, in relation to humans and their myriad roles and in relation to one another. From there

34. Henry Rosemont and Roger T Ames, *The Chinese Classic of Family Reverence: A Philosophical Translation of the Xiaojing* (Honolulu, HI: University of Hawaii Press, 2009), 113–14.

Confucians must consider how universal virtues are to be realized in behavior in each of those roles with their attendant relations. Then they must consider which particular virtues are applicable to each of those roles and how those particular virtues are to be realized in behavior with respect to each other role in the social network. Finally, the Confucian virtue ethicist must consider potential conflicts between universal and particular virtues in generating behavior and between universal virtues and habituated behaviors expressing universal virtues in particular circumstances.

The flexibility advantage in Confucianism clearly comes with a complexity cost whereby the whole framework risks spiraling into an unmanageable chaos. This risk is already potentially there when considering only the many roles with attendant particular and universal virtues to be expressed in behavior in human societies, let alone adding a potential order of magnitude more possible roles for AIs to play. It is not that the Aristotelians do not face a potential complexity crisis as well, but the contemporary revival of virtue ethics, especially as influenced by Alasdair MacIntyre, has sought to manage it by circumscribing the locus of its applicability to small, face-to-face communities.[35] Confucians never attempted such a strategy, having from its inception been a tradition that seeks to shape culture and society broadly from the highest levels, and almost always Confucian thinkers were situated in large, complex societies. Instead, achieving traction on complexity came by enforcing behavioral patterns associated with roles and downplaying individual consideration of whether those patterns in fact accord with either particular or universal virtues, which is to say through the adoption of legalism.[36] Whether or not such a strategy is ultimately adjudicated helpful, healthy, or good, it is nevertheless notable that at least with respect to AIs it is in fact even more easily implemented insofar as such rules governing behavior can be programmed in from the beginning, at least in many instances.

To envision a Confucian virtue ethics analysis of AIs in practice, it is helpful to begin with a relatively prevalent example from the literature, namely self-driving cars and the trolley problem. In this

35. MacIntyre, *After Virtue*.

36. Yuri Pines, 'Legalism in Chinese Philosophy', in *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Winter, 2014, http://plato.stanford.edu/archives/win2014/entries/chinese-legalism/.

thought experiment, a self-driving car suffers a catastrophic failure that results in having to decide between a course of action that results in the death(s) of either the human occupant of the car or a group of people standing along the side of the road. Loss of human life is unavoidable in this scenario, and as originally framed as a challenge for utilitarian analyses, the issue comes down to evaluating the relative value of the individual in the car in comparison with the value of the individuals along the side of the road. Aristotelian virtue ethicists have generally been uninterested in the trolley problem due to their rejection of universal norms, rules, and principles that would enable articulation of a singular, concrete resolution.[37] If virtue ethics are unable to grapple with it, however, then it is unclear that a virtue ethics approach to AI is viable since this is a real, practical moral problem faced in AI development rather than just an ethical thought experiment.

While impossible to give a full analysis here of the problem from a Confucian virtue ethics perspective, it is possible to chart the terrain such a procedure would need to follow. The first step is to understand the role of self-driving cars, which includes but is not reducible to their functional role of conveying people and cargo from one place to another through a range of dynamic circumstances including obstacles. The role of the car is not reducible to this function, though, precisely because other relational aspects serve to further constrain that role. For example, the car may be the property of the person being conveyed, in which case it might have particular role-based requirements of deference to that person, as opposed to being owned by a public entity in which case that deference might be balanced by an equivalent deference to the individuals on the side of the road. The analysis here includes, in part, whether the car is an inferior, equal, or superior to the occupant and each of the individuals it might hit when the catastrophe strikes. The invocation of deference, of course, has already invoked a virtue that is particular to certain roles and not others, and would only be applicable in the case that the car is understood to be an inferior or perhaps an equal.

37. Liezl van Zyl, 'Virtue Ethics and the Trolley Problem', in *The Trolley Problem*, edited by Hallvard Lillehammer, Classic Philosophical Arguments (Cambridge: Cambridge University Press, 2023), 116–33, https://doi.org/10.1017/9781009255615.008.

Once the role is understood, the next step in the analysis is to consider how universal virtues, such as humaneness, apply given the configuration of the roles of the various agents in play. Humaneness is particularly interesting to consider in this case because it is not only universal but synonymous with virtue in general. The result is that if humaneness is the only virtue available upon which to base the required decision in the self-driving car version of the trolley problem, then its generality provides very little traction such that the analysis is dominated by the various roles in play and quickly elides back into a utilitarian analysis of the relative value of each of the people.

Also involved in the analysis would be the role of the occupant vis-à-vis each of the individuals along the side of the road. It may be that the occupant of the car has a particular virtue of loyalty with respect to one or more of the people on the roadside. Even though the occupant is not the moral agent in the case of a self-driving car, the car might take that loyalty into account in its moral decision making. In this sense, the particular virtue of loyalty between the occupant and one or more bystanders is being treated also as a duty, at least from the perspective of the car in formulating its own virtue analysis.

A final point to raise regarding this example for now is that the baseline habituated behavior of the car in its role is likely to be that it should, except in exceptional circumstances, remain on the road, preferably in its lane or at least a lane. A catastrophe is clearly an exceptional circumstance, but a Confucian virtue analysis must find ways to adjudicate whether it is sufficiently exceptional to justify modifying course from the habituated behavior, which is to say the 禮 (*li*; ritual) that governs self-driving cars. Confucians heavily influenced by legalism would tend to the internalist interpretation Ramsey describes and hew closely to the ritual norms. Absent that tendency toward legalism in the tradition, the virtue of humaneness would seem to justify granting exceptionality to a wider range of situations such as the self-driving car catastrophe, and likely many less dramatic interventions.

Unlike self-driving cars, Large Language Models (LLMs) such as ChatGPT are forms of AI designed to interact with humans through the medium of language. The phenomenon of AI hallucinations is when the model generates outputs that are incorrect, impossible, or not based on the inputs. 'The term "hallucination" is used to draw parallels between these unexpected AI outputs and the human

experience of perceiving things that are not actually present in reality.'[38] The model may nevertheless present the outputs as authoritative and correct, and when received by an unsuspecting or insufficiently informed audience these hallucinations may be taken as true. From an Aristotelian virtue ethics perspective, a solution to this problem might be to program the virtue of humility into the model such that the model recursively checks, double checks, and otherwise verifies its results, and presents them less decisively. A Confucian virtue ethicist, by contrast, might question whether humility is a virtue appropriate to the role of an LLM. If, for example, the role of an LLM is to serve as a sort of research assistant, it may be that the relationship with the researcher places the onus more on the researcher to be skeptical of all results at baseline and to have a sufficient level of knowledge of the field to recognize a hallucination when it manifests. In this way, the expectation is that the human demonstrates appropriate intellectual virtues rather than expecting that an AI demonstrate appropriate moral virtues that may not be technically possible, at least as yet.

Finally, the example of AI neural implants highlights the utility of the Confucian focus on roles in virtue ethics and provides a helpful transition to the final section looking at the broader philosophical anthropology in which virtue ethics are framed. While 'devices that interface with the neural system are currently in use and development only for those with a therapeutic need', 'one future use of brain chip implants could be to augment brain functioning for people even without therapeutic need.'[39] Both possibilities give rise to numerous questions about the role of the AI, the role of the host in which the AI is implanted, and how they relate to one another and to others in wider society. Assuming that two people are social equals, does one who then has an AI implanted remain an equal with the person who does not have the enhancement, become their superior, or in fact become their inferior? The possibility of AI neural implants overriding the subjectivity and control of their hosts has become fertile ground

38. Ian Cunningham, 'AI Hallucinations: The Hidden Risks of Machine Learning', in *AI Pathway* (blog), May 11, 2023, https://www.aipathway.com/ai-hallucinations/.
39. Lee Rainie *et al*, 'AI and Human Enhancement: Americans' Openness Is Tempered by a Range of Concerns', Internet, Science, & Tech (Washington, DC: Pew Research Center, March 17, 2022), https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns/.

for a whole genre of dystopian literature and other media. Yet even apart from fears of such imaginings coming to life, the ability to meld human and artificial intelligences provokes a whole set of questions about how roles might change as a result that will require extending the Confucian approach to virtue ethics, especially its consideration of roles. Confucians have already given careful consideration to role alignment across the various levels and sectors of complex societies as evidenced by the extension of the concept of filiality beyond the family to state social systems as described above. The task now is to extend the conception of role alignment to include not only nonhuman moral agents independently but also nonhuman moral agents interfacing directly with the agency of human moral agents. While such a constructive enterprise is beyond the scope of this paper, it is important to emphasize that without the notion of roles it is difficult to see how virtue would apply at this interface, applying to two moral agents independently and as interfaced simultaneously.

AIs and Anthropology

There is a distinct divergence between Aristotelian and Confucian virtue ethics at the level of the philosophical anthropologies framing their respective enterprises and forming the basis for determining what constitutes a moral agent. This divergence also has important implications for how Confucians and Aristotelians interpret the potential for artificial and human intelligences interfacing through AI neural implants. Ultimately, the philosophical anthropology undergirding the Aristotelian project as inherited by the three Abrahamic theistic traditions results in much greater skepticism toward human and artificial intelligences interfacing, whereas Confucians are able to much more easily embrace the possibilities afforded by this prospect.

For Aristotle himself, being human means uniting an animal body with a rational soul, the latter being that which enables humans to accord with virtue in order to achieve our ultimate good, that is, happiness or wellbeing.[40] Presumably, a contemporary inheritor of this conception of human nature could be open to the idea of humans interfacing with AIs through neural implants on the basis that the

40. Aristotle, *Aristotle*, 1097b22–1098a20; Kraut, 'Aristotle's Ethics', sec. 2.

goal of such implants would be to enhance the capacity of reason such that accordance with virtue is likewise enhanced. However, before arriving at such an affirmation, this Aristotelian would have to overcome a degree of skepticism rooted in concern that the interface might interfere with extant rational capacities resulting in discord with virtue.

The situation is quite different for the theistic inheritors of Aristotle, for whom the human capacity for reason, and for Aquinas the natural law, are termini of divine creation which have been deemed good by God. For Jews and Christians particularly, for whom humans are understood to be made in the 'image and likeness of God (*imago dei*)', it is not clear that it is possible to achieve any higher goodness than that with which God has already endowed the capacity to achieve in humans in the very act of creation. Moreover, the risk of meddling in divinely gifted human nature such that a person might no longer reflect divinity is too great. Human nature is understood as relatively fixed as divinely created, and so attempts to change it, through AI neural interfaces or otherwise, would be immoral because in doing so humans are 'playing God'.

Confucianism does not share this concern with humans taking action to change our own nature. In fact, 性 (*xing*; human nature) is merely what we are born with, and the whole purpose of the extensive Confucian tradition of ritual, moral, and intellectual training is in fact to refine, shape, and alter what we are born with so as to become a 君子 (*junzi*; noble person) who embodies 仁 (*ren*; humaneness). This is why Confucianism is known as a tradition of moral self-cultivation:[41] changing our human nature as it is endowed at birth is the whole purpose of the tradition. If AI neural implants can hold out the promise of achieving an even higher degree of humaneness than what humans can achieve with only our meat brains, this would be an exciting potential for Confucians.

This points to a final contrast between the Confucian and Aristotelian approaches to virtue ethics having to do with the conception of the ultimate goal of virtuous conduct: *eudaimonia* (happiness) for Aristotelians and 仁 (*ren*; humaneness) for Confucians. For Aristotelians, happiness is the ultimate goal and is a fulness unto itself of fixed dimension. This is to say that there is no such thing as getting beyond or above happiness, or of extending

41. PJ Ivanhoe, *Confucian Moral Self Cultivation* (Hackett Publishing, 2000).

happiness to greater degrees. It is an end, a cap, a finishing point, the highest good.[42] Not so for the Confucians, who are not so much concerned about humaneness as an end but rather the roots, context, and means of cultivating it. Mencius identifies the root of humaneness in 惻隱 (*ceyin*; compassionate disposition), and all of the early Confucian thinkers emphasize the necessity of it being nurtured, with an elaborate program of formation in ritual and classical texts envisioned by Xunzi.[43] This shift in focus results in a conception of humaneness that is virtually limitless: more training and nurturing, which is to say more moral self-cultivation, will lead to more and more humaneness. Humananess is virtually infinite, rather than a discrete, finite good as Aristotle conceives it.

Conclusion

The conclusion to be drawn is that the Confucian approach to virtue ethics has a higher degree of tolerance for the idea that AIs may be able to play a role in improving humanity, our virtue, and our goodness. Insofar as US citizens reflect these Aristotelian impulses, a recent study from Pew Research Center shows that a majority in the US think neural implants would be bad for society, and seventy-eight percent would not want such an implant, with these views skewing toward those with high as opposed to low religious commitment.[44] China, on the other hand, is heavily influence by Confucianism, and shows greater openness to cloning, gene editing, and pursuing neural implant technologies.[45] Some caution is warranted for the Confucians as well, though, given that while great progress is being made in terms of the interface aspect of neural implants, the AI that will eventually be interfaced still has a long way to go. Increased humaneness may not be the result of implanting an AI that is 'at once both smarter and dumber than any person you've ever met'.

42. Kraut, 'Aristotle's Ethics', sec. 2.
43. Mark Csikszentmihalyi, 'Confucius', in *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Summer 2020 (Metaphysics Research Lab, Stanford University, 2020), sec. 4, https://plato.stanford.edu/archives/sum2020/entries/confucius/.
44. Rainie *et al*, 'AI and Human Enhancement', 83–86.
45. Dennis Normile, 'CRISPR Bombshell: Chinese Researcher Claims to Have Created Gene-Edited Twins', in *ScienceInsider*, November 26, 2018, https://www.science.org/content/article/crispr-bombshell-chinese-researcher-claims-have-created-gene-edited-twins.