Φιλοσοφια
Volume 27, 2: 2026

CONFUCIAN ETHICS AND AI ALIGNMENT

Ranie B. Villaver
University of San Carlos, Philippines

The problem of AI alignment or AI value alignment is the problem of identifying which human value, principle, or ethics is the best with which Artificial Intelligence and Autonomous Systems (i.e., robots) should be designed. Among those that have been proposed is the ethics of Kongzi 孔子 (Confucius) or Confucianism, a fundamentally skills-based ethic. Support for the suggestion of having Confucianism as the best theory, however, has not been fully articulated. In this paper, I argue that support for the suggestion is palpable in three articles on philosophy and ethics of technology and Confucianism, that of Pak-Hak Wong, and Shan Jing and Neelke Doorn. In this paper, I discuss each of these articles and show how they stress Confucian moral excellence, which is the reason why they support the suggestion. My discussion is in this paper's final section. In the first section, I discuss AI alignment, and in the next, I discuss Confucian moral excellence vis-à-vis Buddhist conception, because moral excellence is also the reason that has been given for proposing Buddhism as one rival theory, and because these traditions' conceptions of a morally excellent person are comparable.

Keywords: AI value alignment problem, Buddhism, Confucius's philosophy, moral excellence

INTRODUCTION

Confucian ethics, or the ethics of Kongzi 孔子 (Confucius), a 6th century B.C.E. Chinese philosopher,[1] is a skills-based ethic (Lai 2006). It is about skills, since it conceives of moral deliberation as fundamentally involving the cultivation of reasoning skills and sensitivity. As a skills-based ethic, Confucian ethics can be *sui generis*. It is not like any one of the Western ethical theories, even if concerns about virtuous character, care, duty or obligation, and consideration of consequences are palpable in texts associated with it (Lai 2006, 109-24).[2,3] Might such an ethic be the solution to the problem about artificial intelligence (AI) value alignment? This is a question about the possibility of the ethics as *the best* theory for AI and robotics. In a recent conference held to "discuss particular social and ethical challenges facing Asia with AI" (Checketts et al. 2025, 1), it has been suggested that Confucianism might be that theory. The answer, then, to the question is in the affirmative. The discussion report of the conference says, "…in lieu of only pointing out the dangers and

divergences between Western and Asian ethical approaches to AI, it should be noted that an alternative, authentically Asian approach to AI should be outlined. Confucianism offers many suggestions for what an AI could be." (Checketts et al. 2025, 5). In particular, the report points out that the Confucian value of *xin* 信 (trustworthiness) should be used "in designing the values of the AI" (Checketts et al. 2025, 5). The ground for the suggestion of having Confucianism as *the* solution is not given in the report, although it seems very clear that the reason is that the value of *xin* is rather foremost. In this paper, I argue that support for the suggestion is palpable in three articles on *philosophy* and *ethics* of technology and Confucianism. These articles are: Pak-Hang Wong's "Dao, Harmony and Personhood" (2012) and "Why Confucianism Matters for the Ethics of Technology" (2020), and Shan Jing and Neelke Doorn's "Engineers' Moral Responsibility: A Confucian Perspective" (2020). They stress (Confucian) moral excellence. That they do is the reason why they support the suggestion, and note should be made that moral excellence is the reason that has been given for proposing another Asian religio-philosophical tradition, viz. Buddhism, as the best theory.[4]

I first discuss the alignment problem here. In the next section, I discuss Confucian and Buddhist moral excellence and compare the conceptions. This is because of two reasons: (1) the emphasis on moral excellence is ultimately the reason for the claim that Confucian ethics is the best theory, and (2) moral excellence is also the reason that has been given for proposing Buddhism as a rival theory.[5] In the final section, I discuss each of the articles in question in this paper, and show how they emphasize Confucian moral excellence, thereby supporting the proposal.

AI ALIGNMENT

The problem of AI alignment is the problem of *value* alignment. Artificial intelligence (AI) is "the possession of intelligence, or the exercise of thought, by machines such as computers" (Hauser n.d.). For example, IBM's Watson, which defeated Ken Jennings and Brad Rutter at the quiz show *Jeopardy!*, and now OpenAI's ChatGPT and Google's Gemini (among others), are AI systems (O'Leary 2023, 282-295). They are AI, because they are 'machines' which exhibit some sort of intelligence and somehow appear to possess it. In the *Jeopardy!* game, it takes the ability, among others, to read (and understand) the question and retrieve or think from memory the correct or possible correct answer, or hazard a guess, all of which are intelligent abilities. If IBM's Watson could do those (or appear to do those) and, for that matter, could win the final *Jeopardy!*, it appears we cannot say that it does not have intelligence. The problem of aligning AI with human values and principles appears very lucid in the attempts of organizations or institutions to draft guidelines on the construction of ethically-aligned designs. One such document is that of the Institute of Electrical and Electronics Engineers (IEEE). The document is entitled *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Based on this, the problem of alignment is the task of having human welfare as the foremost consideration in designing AI and autonomous systems (AS) or robots. In their discussion of the first version of *Ethically Aligned Design*,

Chatila et al. (2017) introduce the problem being addressed by the document in the following words: "One can consider [the] examples of AI and autonomous systems (AI/AS) as great achievements or claim that they are endangering human freedom and dignity. We need to make sure that these technologies are aligned to humans in terms of our moral values and ethical principles to fully benefit from the potential of them" (Chatila et al. 2017, 110). According to the words, it is possible that the systems are not created or designed to serve humans or to advance us, and so it is imperative that AI and AS be designed according to the interests of humans. This focus on human welfare and values or principles is the concern of value alignment. Given the foregoing formulation, the question of whether it is possible for Confucian ethics to be the solution to the problem about AI value alignment is the question of whether the ethics is the best theory or the norm that the systems should be designed, i.e., aligned with.

A discussion of an objection to the task of AI value alignment just described is in order. According to the meaning of value alignment, the task involves discovering the best or true theory and then using it as the basis. The objection goes that this is not the task. Iason Gabriel (2020, 425) points out two reasons for the objection. First is that the idea that the essence of morality could be had ought to be seen as a chimera, as "each of the major candidates, at least within Western philosophical traditions, has strongly counterintuitive moral implications in some known situations, or else is significantly underdetermined" (Gabriel 2020, 425).[6] That these theories are "significantly underdetermined" means that the evidence available for judging them objectively true is insufficient. The other reason is "principled disagreement about how best to live" (Gabriel 2020, 425). Gabriel writes the following:

> …even if [the first objection] were not the case and we came to have great confidence in the truth of a single moral theory, this approach immediately encounters a second problem, namely that there would still be no way of reliably communicating this truth to others. For, as the philosopher [John] Rawls [b. 1921, d. 2002] notes, human beings hold a variety of reasonable but contrasting beliefs about value. What follows from the 'fact of reasonable pluralism' is that even if we strongly believe we have discovered the truth about morality, it remains unlikely that we could persuade other people of this truth using evidence and reason alone. (Gabriel 2020, 425)

The point is that the beloved true moral theory would only end up receiving refutations, suggesting, if not indicating, that the pluralistic approach is the way to go.

A response to the challenge of the pluralistic approach of Iason here is the argument that, because all peoples now, amidst the challenge of the powerful technologies, should have guidelines or soft laws, each should remodel their own cultural or religious traditions, for the simple reason that we might not be amenable to the values currently being seen as prominent for AI adoption or rejection. In "Non-Western AI Ethics Guidelines: Implications for Intercultural Ethics of Technology," Soraj Hongladarom and Jerd Bandasak (2024) defend the belief that the various ethics guidelines for AI and robotics of nation-states cannot be not informed by their own unique empirical conditions and cultural traditions for them to contribute to the task of

carrying out "a strong international collaboration on AI ethics" (Hongladarom and Bandasak 2024, 2031). They point out that it is the "clash with emerging powerful technologies" that has challenged non-Western nation-states (viz. Thailand, China, Japan, Korea, Dubai, and India) to come up with AI ethics guidelines. They point out that several principles of the guidelines appear to reverberate those in Western formulations. That they do, according to Hongladarom and Bandasak, appears to imply that *ethical universalism* has triumphed over *particularism*. They, however, argue that the appearance is not reality or factual. According to their analysis, it is clear that the presence of culture-specific concerns is borne out by the surveyed non-Western guidelines (most notably by those of Thailand, China, and Japan). Hongladarom and Bandasak express the worry about the absence of *explicit* cultural considerations in the guidelines, for if the nation-state guidelines neglect cultural roots and adopt only foreign principles, "it is rather difficult to see how [the] principles [and thus the guidelines themselves] can be implemented in any detail in practice" (Hongladarom and Bandasak 2024, 2027). It is to be conceded that the guidelines are attempts by people of the world to accommodate "the empirical conditions of today's globalized world" (Hongladarom and Bandasak 2024, 2020). Their suggestion (for philosophers in Asia) is to think about ethics or morality *à la* the Prussian philosopher Immanuel Kant (b. 1724, d.1804): just as Kant "remodel[ed] Protestant morality so that it appeared modern, i.e., without an explicit reference to God" (Hongladarom and Bandasak 2024, 2028), thinkers in this part of the world should also *remodel* their own cultural or religious traditions. Hongladarom and Bandasak's idea that people should *remodel* their tradition means that each must think that the values each holds are not irrelevant to facing the challenges of the present. This means that it is still possible to think of the best theory for AI and robotics and to suggest it.

MORAL EXCELLENCE IN CONFUCIANISM VIS-À-VIS BUDDHIST CONCEPTION

The point that the Buddhist theory's conception of a morally excellent person (whose one main feature is that of compassion) appears clearly to be comparable to Confucian *junzi* 君子 (paradigmatic individual) can be a reason for saying that indeed it is possible to have Confucianism as *the* theory for AI and robotics. This is because Buddhism has been proposed as that theory. In this section, I discuss this proposal briefly first. This section's bulk is a discussion of moral excellence in each of the traditions. Soraj Hongladarom's *The Ethics of AI and Robotics: A Buddhist Viewpoint* (2020) is one such work that endorses Buddhism as the theory for AI. The book is Hongladarom's contribution to the task of identifying elements or points in the Buddhist teaching that appear to have something to say about the powerful technologies of AI and robotics. Essentially, Hongladarom (2020, 4) points out that his book elaborates on a Buddhist solution to "the problem of how to think about AI ethics most effectively." He is proposing that the Buddhist perspective he discusses is that theory which is the "best one for AI ethics" (Hongladarom 2020, 4).

Moral excellence in Buddhism is clear in the theory's concepts of *arahant* and *nibbāna*. The *arahant* is a being who has attained *nibbāna* or enlightenment

(Hongladarom 2020, 33). Because Hongladarom's (2020, 8) discussion of machine enlightenment (which is pivotal in his endorsement of Buddhism as the best theory) hinges on *arahant*, the enlightened being, his discussion is important here. Hongladarom's discussion of the Buddhist perspective fundamentally centers on *nibbāna*. Since *nibbāna* attainment essentially involves overcoming the belief in the existence of a permanent and distinct self, an enlightened being is one who "completely integrates himself or herself with nature so that there is no separation between that self and what is outside [everything surrounding, i.e., nature]" (Hongladarom 2020, 79). Hongladarom claims that there can be *machine enlightenment*, i.e., machine achievement of the ethical perfection characterized by consideration of other beings, environment, or nature. Enlightened artificial general intelligence or AGI (or strong AI) would show consideration and prioritization of others' interests more than its own. Superintelligent AGI or robots cannot be conceived of as not consisting of or as not having such ethical wisdom, because superintelligence entails full awareness of the truth of no-dichotomy of self and nature and hence absence of or the elimination of ignorance, *avijjā* (Hongladarom 2020, 78-83). There can be enlightened artificial specialized intelligence or ASI (or weak or narrow AI), on the other hand, because it would be based on the manufacturer's design.[7] According to Hongladarom, in that *nibbāna* (and machine enlightenment) is attainment of ethical perfection, the Buddhist teaching of moral excellence elicits the idea of both *technical* goodness and *ethical* goodness. In Buddhist teaching, *technical* goodness and *ethical* goodness are interlinked, since following the rules for self-cultivation (*sīla*) is the technical matter for achieving moral excellence (Hongladarom 2020, 6, 86-9). Hongladarom's conception of *a good* AI comes from this. A *good* AI "embodies" not just technical goodness but also ethical excellence. Hongladarom's (2020, 5) account uses a car analogy, the idea of which is that just as a good car is one that not only meets the technical requirements but should also have the feature of *safety*, a good AI should have the feature that ensures the safety or well-being of all concerned. This feature is understood by Hongladarom as the *consideration and prioritization of others' interests more than its (or its manufacturer's) own* (Hongladarom 2020, 5-6). More specifically, he understands it to mean as the stipulation that AI "must find ways to relieve all others, as much as it possibly can, of suffering" (Hongladarom 2020, 6). Machine enlightenment is Hongladarom's response to the problem of AI alignment.[8]

Moral excellence in Confucianism is clear in discussions of the Confucian *junzi*. Here, it must be first noted that I am using the translation "the paradigmatic person" for 君子 (*junzi*).[9] Although "noble person" or "gentleman" also conveys the idea that the *junzi* is a model or a role model, it is "the paradigmatic person" which is straightforward. Here, I will use Antonio S. Cua's (1971) elucidation of *junzi* as an important concept in Confucian philosophy. Cua's discussion is found in his elaboration on Confucius' ethics of flexibility. Confucius' ethics is one of *flexibility* and *adaptability*, because it is about both *jing* 經 and *quan* 權. The doctrine of *jing-quan* is the doctrine of "the *normal* and the *exigent*, or the normal and the exceptional" (Cua 1971, 51), the idea of which is that norms (*jing*) and the 'weighing' (*quan*) are important (Lai 2012, 350). Cua presents and discusses two passages in the *Mencius*

(1B8 and 4A17) to illustrate the doctrine (Cua 1971, 51-2). The passages are as follows:

> King [Xuan] of [Qi] asked, 'Is it true that [Tang] banished [Jie] and King Wu marched against [Zhou 紂]?'
>
> 'It is so recorded,' answered Mencius.
>
> 'Is regicide [then] permissible?'
>
> 'He who mutilates benevolence [*ren* 仁] is [called] a mutilator [*wuzhizei* 謂之賊]; he who cripples rightness [*yi* 義] is [called] a crippler [*wuzhican* 謂之殘]; and a man who is both a mutilator and a crippler is [called] an "outcast" [*wuzhiyifu* 謂之一夫]. I have indeed heard of the punishment [*zhu* 誅] of the "outcast [*yifu* 一夫] Zhou [紂]" but I have not heard of any regicide [*shijun* 弑君].' (Lau 2003, 23)(2.8/11/9-16)[10]

> [Chunyu Kun] said, 'Is it prescribed by the rites [*li* 禮] that, in giving and receiving, man [男] and woman [女] should not touch each other [授受不親]?'
>
> 'It is,' said Mencius.
>
> 'When one's sister-in-law is drowning, does one stretch out a hand [手] to help her?'
>
> 'Not to help a sister-in-law who is drowning is [是] to be a brute [豺狼]. It is prescribed by the rites that, in giving and receiving, man and woman should not touch each other, but in stretching out a helping hand [手] to the drowning sister-in-law one uses one's discretion [*quan* 權].'
>
> 'Now the empire [天下] is drowning. Why do you not help it?
>
> 'When the empire is drowning, one helps it with the Way [道]; when a sister-in-law is drowning, one helps her with one's hand. Would you have me help the Empire with my hand.' (Lau 2003, 84)(7.17/38/20-30)

The first passage, 1B8, points out that exterminating a tyrant ruler in the person of Zhou 紂 would not count as *shijun* 弑君, literally "kill a prince-ruler", or regicide. It would count only as "executing an outcast" [*zhuyifu* 誅一夫], because he was a crippler [*can* 殘] and a mutilator [*zei* 賊] of *ren* 仁 (benevolence) and *yi* 義 (rightness) – and thus was *yifu* 一夫 , an "outcast" or, in my translation, "just a fellow". The second passage, 4A17, points out that the *li* 禮 (rite; rule) of men and women not touching in giving and receiving may be forgone, because "not to help a sister-in-law who is drowning is to be a brute [*chailang* 豺狼]," i.e., to be not human. The reason why the passages are about *jing-quan* is that in both cases there is a clear norm or rule to be applied (rigidity) but the exigency or abnormality of the concrete situation calls for abandonment of the norm (flexibility) (Lai 2012, 350). In the article, Cua points out that this ethics of Confucius is clear in *junzi*. He identified the following features of *junzi* which the *Analects* gives: a person of (1) "moral virtues pervaded by an affectionate concern for [*ren*]," (2) "propriety (*li* 禮) and righteousness [*yi*]," (3) "catholicity and neutrality," and (4) "his words and deeds" (Cua 1971, 42-9).

Concerning the first feature, Cua points out that *junzi* is a person of moral virtues such as generosity, sincerity, kindness, and courage, all of which are "pervaded by the ideal of [*ren*] (human-heartedness)" (Cua 1971, 42). One of the passages which Cua cites here is 4.5, that is, book 4, chapter 5. He quotes in particular the following: "君子去仁, 惡乎成名？君子無終食之間違仁, 造次必於是, 顛沛必於是." (4.5/7/14-15),[11] translated by Simon Leys as, "If a gentleman [君子] forsakes humanity [仁], how can he make a name for himself? Never for a moment does a gentleman part from humanity; he clings to it through trials, he clings to it through tribulations" (Leys 1997, 15). This quoted section appears clearly to culminate the chapter/passage which begins with the Master's (Kongzi's) point that it is not proper for a person to go against one's principles (*dao* 道) in order to obtain riches and fame or to escape destitution. The passage points out that the *dao* is *ren* 仁. Accordingly, *junzi*'s *dao* is *ren*. This, in turn, clearly appears to mean that *ren defines junzi*, which is Cua's point.

About *junzi* being a person of *li*[12] and *yi,* Cua points out connections of the two concepts with morality that is in *ren*: "If *li* is the emphasis on the contact between [*ren*]-morality and the cultural lifestyle, [*yi*] is that on the contact between [*ren*]-morality and actual situations" (Cua 1971, 45). By cultural lifestyle, Cua means "the form and possibility of moral achievement within the cultural setting" (Cua 1971, 44). To support this meaning of *li*, Cua cites 6.27: "子曰：'君子博學於文, 約之以禮, 亦可以弗畔矣夫!'" (6.27/14/10), which Leys translates as: "The Master said: 'A gentleman enlarges his learning (*xue* 學) through literature (*wen* 文) and restrains himself with ritual (*li* 禮); therefore, he is not likely to go wrong" (Leys 1997, 15). Since *wen* 文 here is *writing* (or "literature") or *culture*, the idea is that *li* does not pass over culture or "cultural lifestyle". Cua cites 15.18, to point out the relation between *li* and *yi*, even though, according to Cua, the latter is not elucidated in the *Analects*. The passage is as follows: "君子義以為質, 禮以行之, 孫以出之, 信以成之. 君子哉!" (15.18/43/11), which Leys translates as, "A gentleman takes justice [*yi* 義] as his basis [*zhi* 質], enacts it in conformity with the ritual [禮], expounds it with modesty [*sun* 孫], and through good faith [*xin* 信], brings it to fruition. This is how a gentleman proceeds" (Leys 1997, 76). To Cua, the passage shows the relation between the concepts because it points out that it is *yi* or "moral action" which takes precedence and yet the action departs not from *li* (Cua 1971, 45).

Concerning catholicity and neutrality in *junzi*, Cua discusses the paradigmatic individual's non-committal to specific courses of action. It is the problem of neutrality and paradigm which Cua focuses on in his paper's second and final sections. According to Cua, the problem is clear in *Analects* 4.10: "子曰：'君子之於天下也, 無適也, 無莫也, 義之與比.'" (4.10/7/27). This is translated by Leys as: "The Master said: 'In the affairs of the world [*tianxia* 天下], a gentleman [君子] has no parti pris; he takes the side of justice [*yi* 義]" (Leys 1997, 16). The problem of neutrality and paradigm is that it is true, according to the passage, that *junzi* bases his actions not on prejudice but on *yi* (which makes him free to a certain extent only, due to *li*-observance), yet how is such freedom or neutrality that which is paradigmatic in *junzi*? Cua states that it is rather the *making of ruling* which is paradigmatic. He writes:

> … [body of rules are taught] but in the dynamic situations of human life we need to make rulings even in the absence of given rules. This is the logic of [*junzi*]. ... He is a paradigmatic guide for ordinary moral agents by virtue of his ability to cope with the changing circumstance within the Confucian moral point of view. By his neutrality of attitude toward specific courses of action, he preserves his freedom of action. The significance of the Confucian notion of [*junzi*] thus lies in its suggestion of a conception of a reasonable moral agent who lives within a common form of life. (Cua 1971, 52-3)

According to these words, the neutrality only indicates *junzi*'s freedom; it is not what makes him paradigmatic. According to the words, what makes *junzi* a paragon is, to use the words of Karyn Lai in spelling out *quan*'s (權) meaning in the *Shuowen jiezi* 說文解字, "the acumen … to know when it is appropriate *not* to follow the norm" (Lai 2012, 350). Accordingly, neutrality is a stance towards the courses of action, and it is *the process of deliberation* that *junzi* carries out, which is paradigmatic. But in that *junzi* is a person of *ren*, *li*, and *yi*, attention must ultimately be given to *ren*-cultivation or self-cultivation. This, in other words, means it is not incorrect to say that it is *knowledge-how* (rather than *knowledge-that*) that is in Confucius, but it is, more precisely, *knowing-to* (act in the moment) (Lai 2012). *Knowing-to* is different from *knowledge-how* because *knowing-to* involves correct, deep appreciation on the part of the agent of the subject matter at hand. In Confucian terms, this means that *knowing-to* involves self-cultivation or self-realization – there is "centrality of character… the fundamental concern is not with information or isolated actions but rather with actions that are effected *as manifestations of* an exemplary life" (Lai 2012, 359). It is Joel Kupperman's *style* that comes close to this point of Lai here. I elaborate on Kupperman's *style* in the discussion of Pak-hang Wong's paper (2020) in the next section.

On the fourth feature, Cua writes that it is about "suiting one's words to the action" (Cua 1971, 48). I suggest that this feature is also seen through a passage in the *Mencius* (*Mengzi* 孟子), passage 3B8. Since 3B8 conveys moral sensitivity as the *junzi*'s *dao* (君子之道 *junzi zhi dao*), it also shows the first feature of *junzi* as a person whose good qualities are all about *ren* (human-heartedness). The passage is as follows:

> When Mencius was in Song, Dai Yingzhi said to him, "At the present time, we would not be able to abolish our customs and market taxes and rely solely on a tax of one in ten. How would it be if we simply reduce the current rates a bit and repeal those other taxes next year?"
>
> Mencius said, "Let's say there were a man who stole one of his neighbor's chickens every day. If someone said to him, 'This is not how a *junzi* behaves!' and he replied, 'How would it be if I simply reduce my current rate a bit, steal a chicken every month instead, and then stop next year?
>
> "If you realize it is wrong, stop immediately. Why wait for next year?" (Eno 2016, 78)(6.8/34/7-11)

Mencius, in the passage, seems to clearly hint that stealing is not *junzi*'s *dao*. This is seen in the sentence (是非君子之道 *shifeijunzizhidao* [6.8/34/10]), which could also be translated as "Such is not the way of the paradigmatic person!" Mencius's answer to the question of whether Dai Yingzhi's proposal is acceptable is that it is not. It is not acceptable because the way of the paradigmatic person (or the process of moral self-cultivation) is serious. It is not lax. Accordingly, to suggest that the passage's message is that there is urgency of ceasing the doing of wrongful action or not starting it is not incorrect. That is because to say that the *junzi*'s way is not lax means that doing what is right cannot be postponed. This interpretation has been put forth by James Legge (1930, 670) and Bryan Van Norden (2008, 83). Just as students would not take a course requirement (say, an essay) seriously if the course convenor keeps on changing the requirement's due date and there is no clear or set marking criteria, the would-be *junzi* would not take righteousness seriously if it is not serious, if righteousness is not urgent. Pointing out that it is urgent to cease doing a wrongful action or that it should not be done (or begun) is achieved by Mencius through an analogy. In the passage, Mencius likens Dai Yingzhi's proposal to stop ridiculous, harmful taxes after a year of reducing the rates to that of a man who steals a chicken every day from his neighbor. To Mencius, the answer to the question of whether the halting of the wrongful action is urgent is 'Yes'.

If that is the passage's message, then there is the question of why Mencius recommends such. Could it be that the answer is Mencius's view, in *Mencius* 2A6, that humans have *xin* (heart-mind) "which cannot bear to see the sufferings of others" (Legge 1930, 548)? Dai Yingzhi's ridiculous tax practices harm the people and the stealing harms the neighbor (if not immediately, then in the long run). Might it be that because wrongful actions make others suffer and we have hearts that "cannot bear to see the sufferings of others," Mencius thinks ceasing the doing of wrongful actions is urgent? It seems that the urgency is due to compassion. Mencius's thought-experiment of the child about to fall into a well is used to show that humans have hearts that are not unfeeling towards others. According to Van Norden, the experiment shows that Mencius's only claim is that "any human would have at least a momentary feeling of genuine compassion, and that the reaction would occur suddenly" (Van Norden 2008, 46).

Be that as it may, it appears that the compassion here conveys an imperative to do something to save the child. In other words, the child is about to experience suffering, and our hearts urge strongly that we must save her and that saving her ought to be done right away. There is urgency. Compassion is the reason for the urgency. Additionally, there is urgency because it appears that there is a mismatch between words and deeds. The idea appears to be that it cannot be that if one's judgment concerning an action is that it is not benevolent, delaying its end or halt is the correct way to go. That just shows one's decisions do not suit one's pronouncements. If compassion is the reason for the urgency, then 3B8 of the *Mengzi* text is about moral sensitivity. The feeling of commiseration or compassion that Mencius demonstrates that we all have appears equivalent to 'regard for the welfare of the other' or sensitivity to the other. According to *Mencius* 3B8, that others have needs is not neglected by a *junzi*. The people of Dai Yingzhi and the man's neighbor have needs. The people find

the taxation burdensome, and the neighbor needs the chickens perhaps for food or, perhaps, for commerce (making a living). Sensitivity to others is sensitivity to their needs. This is moral sensitivity. This analysis of the passage reveals Mencius's view that *junzi*'s *dao* includes moral sensitivity to the other. Moral sensitivity to the other is sensitivity to the needs of the other, a disposition synonymous with 'regard to the welfare of the other.'

Junzi is the exemplar of moral excellence in Confucianism. *Junzi* is the sage in Stoicism and the *arahant* in Buddhism. The traditions of Buddhism, Stoicism, and Confucianism converge. Hongladarom (2020, 8) states that the Buddhist viewpoint he offers is not unlike many ancient Greek ethical theories. As he writes: "… the Buddhist ethical theory has many affinities with many Hellenistic, post-Aristotelian theories, most notably Stoicism" (Hongladarom 2020, 8). In recent years, comparative analysis between Confucianism and virtue ethics has been undertaken rather vigorously; it has been suggested that a good, promising comparative analysis be made between Confucianism and virtue ethics as a way of "doing ethics" (Tiwald 2010, 60). It is here that we find affinities between Confucianism and Stoicism as conceived by scholars. Justin Tiwald (2010) cites Stoicism as an example of virtue ethics, which has the sense of "way of 'doing ethics'" (Tiwald 2010, 60-1). According to Tiwald, Stoic virtue ethics is akin to the practice of medicine. In that the sick soul is afflicted with character flaws, the ethical task is to get rid of those flaws to cure the soul. To Tiwald, such characterization of virtue ethics (as self-cultivation focused or orientated) is one point of convergence or affinity between the Stoics and the Confucians. The claim I am making here is that, given the affinity between the Stoics and the Confucians, and the one between the Stoics and the Buddhists (as pointed out by Hongladarom), there is also affinity between the Buddhists and the Confucians. It may be said thus that *junzi* of the Confucians is the *arahant* of the Buddhists. It must be said here, in concluding this section, that, as is clear from the foregoing discussions, the one main characteristic of sensitivity to the other or compassion (in *junzi*) is that which makes the Buddhist *arahant* and Confucian *junzi*, ultimately, comparable or similar. It is the *exemplification of compassion* of each that makes them comparable or similar.[13] It is not the case that what makes them comparable is that each is viewed in the respective philosophical traditions as a moral exemplar, because if this were so, then the moral exemplar in the theory of ethical egoism, for example, would be comparable to an *arahant*. Ethical egoism's exemplar is not comparable to *arahant* because the theory is one that endorses the total disregard for the welfare of others. In Hongladarom's discussion of *nibbāna* (which the *arahant* has attained) and machine enlightenment, as already presented, each leads to both technical excellence and ethical excellence, the latter of which is fundamentally about consideration and prioritization of the interests of others or, in short, *compassion*. Besides, it is clear in the Buddhist *Parable of the Poisoned Arrow* that Buddhist compassion is fundamentally about concern for the welfare of the other (Hongladarom 2020, 39). The parable demonstrates that questions of who shot the arrow, who made the arrow, what the arrow is made of, among others, are strange questions, because the urgent task is to go to the aid of the person who got shot by the arrow.[14]

THE THREE PAPERS

Two of these three papers on *philosophy* and *ethics* of technology and Confucianism are by Pak-Hang Wong: "Dao, Harmony and Personhood: Towards a Confucian Ethics of Technology" (2012) and "Why Confucianism Matters for the Ethics of Technology" (2020). The third paper is Shan Jing and Neelke Doorn's "Engineers' Moral Responsibility: A Confucian Perspective" (2020). It must be noted that Wong is one of the authors of that published conference report (Checketts et al. 2025) and it is in his institution at which the conference was held.[15] In Wong's two papers, I discuss here, there is no mention of Confucianism *and* AI alignment. It is only in the report that this is stated.[16] On page 2 of the report, under machine ethics frameworks (which is about that framework on which to train AI, value alignment, and AI aim or function) the following are given: "Value alignment as well needs to be considered in terms beyond Western discourse. Harmony, propriety, or a strong concept of trustworthiness, for example, are cultural values that East Asian countries may wish to build into their AI systems, which are not prioritized in Western cultures." (Checketts et al. 2025, 2). According to this point, concerning the AI value alignment problem, a cluster of non-western points to think about are those Confucian concepts of *he* 和 (harmony), *li* (behavioral propriety), and *xin* (trustworthiness). These concepts are discussed or are alluded to in the three papers I now discuss in detail here. (It ought to be noted that just because it is only in the conference report that we see Wong's point about the connection between Confucianism and AI alignment, it does not mean that the idea is not "nascent" if not palpable in his two works I discuss here.) At the end of the discussion, I show how the papers stress (Confucian) moral excellence, since it is that emphasis which indicates that they support the thesis that it is Confucian ethics that is the best theory. It is moral excellence (which, as discussed, is essentially characterized by exemplification of compassion) in the philosophy of Confucius that makes the philosophy the best for AI and robotics.

Wong (2012) attempts to articulate a possible convergence between traditional Confucian ethics and *philosophy and ethics of technology.* He discusses important themes in Confucian philosophical ethics and, in light of the themes, considers attitudes we might have to seriously consider amidst technology. Wong discusses the themes of *Dao* 道, harmony (*He* 和), and personhood in Confucian philosophical ethics because they are "*most* relevant for an account of ethics and technology" (Wong 2012, 68), and this is in response to the question of whether it is possible to construct a Confucian perspective on ethics and technology given the enormous field of study covered by the Confucian tradition. Wong ultimately points out "four major considerations that should be characteristic of a Confucian ethics of technology" (Wong 2012, 80-3).

The first pertains to the idea that given that *Dao*-realization is "the ultimate goal and ideal of human beings" (Wong 2012, 80) in Confucian ethics, the tradition's answer to the question of whether it is *the right* or it is *the good* that has the priority is neither. *Dao*-realization has the priority and the good and the right are both under its purview. Wong writes that what this implies is that a Confucian ethics of technology does not favor the turn in philosophy and ethics of technology towards well-being and

the good life. What it endorses is the move to "bring in prudential considerations, i.e., the good, or, to incorporate moral considerations, i.e., the right and the just" (Wong 2012, 80) – although, ultimately, it appears that it is Wong's point that the basis of which is (Confucian) virtue ethics in its radical form.[17]

The second consideration is about the ideal of harmony, which relates to both "*applying* Confucian ethics on issues in ethics and technology" and "with *formulating* issues in ethics and technology from a Confucian perspective" (Wong 2012, 81). Concerning application, the idea is that harmony becomes the end or *standard* – Confucian ethics favors not the turn to the good life but the inclusion of discussion about right and wrong. Concerning formulation, the idea is that treatment of ethical issues in technology would center not on "final answers" but rather on the *focal*, on "know-how and harmony-as-a-skill" (Wong 2012, 82). This is the case, given that these are clear and are front and center in Confucianism. The third consideration is about the significance of social roles, which implies that Confucian ethics of technology would worry about the impact of technology towards human social roles and the corresponding responsibilities. Wong writes that the reason for this is that "technologies can either enhance or deter people's fulfillment of their role responsibility, and they can also change the nature of the social roles" (Wong 2012, 82). The fourth consideration is "closely connected" to the third in that this consideration of significance of practice and lifestyle involves (i) looking upon the actual and potential practices which technology elicits, and (ii) examining the effects of these practices on the carrying out of one's roles in society and hence to lifestyle, "the way one lives" (Wong 2012, 83). Wong mentions work on personal robots and reckons that the view that interactions we have with such technology have become human-like shows implications of the Confucian view on ethics and technology (Wong 2012, 83).[18] The idea is that Confucians would have us think about our relationship with technologies, because they either enhance or impede performance of our roles and our being.

Wong (2020) articulates the implications of *li* (ritual), 禮, in Confucian ethics for ethics of technology – with focus on the concept's *communicative*, *formative*, and *aesthetic* functions. Wong attends to *li* substantially in this paper. He begins with a discussion of the work of Joel Kupperman (b. 1936, d. 2020) on *style*. To Wong, it is Kupperman's view that it is important to have "a close at look people's style of interaction, for it communicates people's attitudes (about others) and shows themselves to others, which are essential in ethically fruitful connections with others" (Wong 2020, 613). Style or naturalness (or harmony) – which is about education or training through *li* (ritual), *refinement* and thus *ethics* (Kupperman 1968) – is important to Kupperman because it shows that Confucian ethics does *not* exclude "big moment ethics" (Kupperman 2002, 50).

"Big moment ethics" is one that emphasizes "major choices at ethical crosswords" (Kupperman 2002, 40) and is thus one which "in effect treats almost all of life apart from the big moments as an ethical free-play zone, in which one can do whatever one likes" – "yield[ing] an ethics that does not make demands at all often, and certainly not continuously" (Kupperman 2002, 40). It is the focus on *li* in Kupperman's style which is important to Wong. Wong then points out the ethical

importance of *li* through its functions. The communicative function is captured in the point of *li* as "a culturally-specific 'body language'" (Wong 2020, 615). Wong writes that in relation to the socio-ethical dimension, this means that *li* would be "a shared resource for understanding and interpreting need and care – or, for that matter, other important shared values as well" (Wong 2020, 616). In other words, *li* is that language available as a tool for understanding one another, without which we would be nobody or nothing.

The formative function pertains to *li*'s both delimiting or regulative and ennobling power pointed out in the *Xunzi*.[19] This means that *li* brings people to "become accustomed to the right emotional responses and behaviors, thereby transforming their dispositions" (Wong 2020, 616). The formative function, in other words, is the function of education. In relation to the socio-ethical dimension, Wong writes that this function is about *li*'s role in the inculcation of "norms and values, and enables them to react ethically to different situations in spontaneity" (Wong 2020, 616). *Li* fulfills this role because the formation or education is with attention to *ren* (benevolence) and *yi* (rightness; appropriateness) (Cua 2002, 481). The aesthetic dimension pertains to one's practice of *li* as "beautif[ying] one's emotional response and behaviors by making them more pleasant and agreeable, thereby reducing the potential for conflict and encouraging social cooperation" (Wong 2020, 617). The beauty in question refers not just to *form* but in a significant way also to *li*'s close connection with respect, kindness, or humanity (*ren*). According to Wong, the first function gives the point that it is necessary to examine (i) "not only *what* values are embedded in technology, but *how* these values are, or can be, manifested through the use of technology and in technologically-mediated interaction" and (ii) "the recipients" and "the existing styles of interaction, social conventions, and manners in a community" (Wong 2020, 618). An "example" Wong gives for this is Kamphof's discussion of caregivers' use of tele-monitoring systems. Wong writes: "Kamphof notices that the caregivers' concern is not only about the privacy – or lack thereof – in the system per se but also about *how* the value of privacy and a good caregiver-patient relation are, or can be, realized in its use with the elderly clients" (Wong 2020, 618). I believe some of the questions here are: *Can there be the value of privacy, given the use of such technology? If there is none, can the caregivers be trusted with the information?*

The formative function implies attending to "the *bodily* and *affective* impacts of technology with reference to *Li* in a community and its tradition. It may even warrant *proactively* shaping individual's bodily and emotional states in accordance with Confucian *Li* through the use of technology and in technologically mediated interaction" (Wong 2020, 620). The example Wong gives here is Kristina Niedderer's Mindful Design, which refers to designing technology with the eye of consideration of the welfare of others (Wong 2020, 620). Indeed, we can imagine a mobile phone that "'shouts back' at its users should they be talking too loudly in public places" (Wong 2020, 620-621). The idea is that given this function, "technology can be 'ritualized' to support people's ethical development" (Wong 2020, 621). The aesthetic function entails intertwining of the aesthetic and the ethical. Wong gives the example of Benjamin Grosser's *Facebook Demetricator*, which is a web browser extension that allows users to hide all the platform's metrics. The idea is that the social media

platform's design (with or without the metrics display) surely has effects on the users' interactions, and so the beauty of the design and its ethical impact are important considerations (Wong 2020, 621-2).

Jing and Doorn (2020) is mainly a discussion of Confucian moral responsibility for engineering ethics and engineering ethics education. In the paper, the Citicorp Building case is considered to see how the Confucian conception might respond to vis-à-vis Western analysis of the case. As a work about exploring the Confucian tradition's possible contribution to engineering ethics, its implication for the ethics of the powerful emerging technologies in question concerns the ethical responsibility of the manufacturer or creator. This is because the article's focus is the Confucian perspective regarding the Citicorp case. Jing and Doorn use the tradition's four kinds of moral responsibility, namely, *self-responsibility*, *family responsibility*, *professional responsibility*, and *responsibility to the universe* (Jing and Doorn 2020, 240-4). These kinds are related, because the lattermost responsibility to the universe depends on the others, with self-responsibility as the foundation, on which family responsibility rests. Engineer William J. LeMessurier, according to Jing and Doorn, was professionally responsible, by taking action upon learning about the edifice's structural flaw, which potentially could cost the lives of 200,000 people because the flaw could lead the building to collapse.

Professional responsibility, in Confucian thought, derives from feelings of *xiu* 羞, shame, and *wu* 惡, dislike (Jing and Doorn 2020, 248), pointed out in the fourth century B.C.E. Confucian thinker Mengzi or Mencius.[20] This is different from Western analysis in that there is an absence of distinction between technical and non-technical excellences (Jing and Doorn 2020, 237). The non-technical excellences are the moral virtues, while the technical ones refer mostly to the virtues relating to expertise or technical know-how. Technical virtues "include those capacities and sensitivities closely related to the technical side of engineering that cannot be expressed well in rules, such as mastery of the relevant aspects of mathematics and physics, engineering science, and design, but also the virtues that allow one to make well-considered decisions in situations of complexity or risk" (Jing and Doorn 2020, 237). The Confucian "edge" is clear in Jing and Doorn's point that while in the Western perspective, LeMessurier may be criticized for "withholding critical information" (as Eugene Kremer has done),[21] in the Confucian perspective, there is no such condemnation because the focus is on "harmony and avoidance of shame" (Jing and Doorn 2020, 248). The point of Jing and Doorn is that by keeping the situation secret, "he was able to avoid damaging his reputation, not only for himself but also for the company" (Jing and Doorn 2020, 248), which means that LeMessurier made sure that there was balance or that he harmonized.[22] Jing and Doorn also highlight the Confucian contribution of family responsibility. The idea is that such responsibility focuses on the personal accountability of LeMessurier with regard to himself and to others, but also of accountability of everyone concerned. As they write: "… from the perspective of Confucianism, LeMessurier not only took responsibility for his own family, but also protected the greater 'family' associated with Citicorp. But other people also have responsibilities. Everyone shares a responsibility to keep societal order intact" (Jing and Doorn 2020, 249).

The Confucian perspective on ethics and technology that Wong has sketched out in "*Dao*, Harmony and Personhood" is "far from complete" (Wong 2012, 83) but, together with his articulation of implications of *li*'s three functions for Confucian ethics of technology in "Why Confucianism Matters for the Ethics of Technology" (2020), the perspective points toward a view of a Confucian solution, that of a view of moral excellence. *Dao*-realization and focus on the ideal of harmony are all about such excellence. The goal of *realizing Dao* involves self-cultivation. Although it is characterized by Wong as "the acquisition of virtues [*de* 德]" (Wong 2012, 71), self-cultivation in the Confucian tradition includes specifically education or formation through *li* (ritual). Wong's articulation of implications of *li*'s three functions do lead to a view of moral excellence, in that the implications are all about ethical development and impact. Jing and Doorn's (2020) discussion of Confucian moral responsibility also leads to the view because the kinds of responsibility underscored are about ethical development and moral impact. I would also like to note that this Confucian view of moral excellence is not unlike that of the Buddhist perspective we see in Hongladarom's (2020). *Li* is like the Buddhist *sīla*, the rules for self-cultivation, in that *li* is "behavioral propriety" or the codes of proper conduct. Ethical excellence or being *junzi* is achieved with, among others, the practice of *li*. In conclusion, in that the three papers do stress Confucian moral excellence they support the proposal to have Confucian ethics as *that* ethics for AI and robotics. The proposal stands on the firm belief that Confucian moral excellence achieved through *li*-practice and exemplified also by *xin* 信 (trustworthiness) and harmony (*he* 和) is important.

CONCLUSION

Before concluding this paper, because in characterizing Confucian ethics I have subscribed to the ethics as one about skills or is skills-based, it is important to address the question of how (i) the three supporting papers' conception of Confucian ethics and (ii) *junzi* as exemplar of moral excellence cohere with it. The three papers' approach to Confucian ethics of technology is arguably not skills-based but based on (social) roles (more palpably in Wong's), and Cua's discussion of *junzi* appears to convey nothing on skills. I discuss the connection of *junzi*'s characteristics discussed by Cua with the analysis of the ethics as skills-based first. That the ethics is skills-based means that ethical decision-making stressed in Confucian moral thinking has the features of "skills of interpretation and creativity, sensitivity to morally significant factors, a broad knowledge and understanding of situations in life, a depth of experience including learning from the experiences of others, and the fine balancing skills required in deliberation and judgment" (Lai 2006, 124). *Junzi*'s features figure in these reasoning elements, and in fact, in Cua's discussion of the problem of neutrality and paradigm, it is claimed that it is the deliberation process of *junzi* that makes him paradigmatic. For example, both Cua (1971, 52) and Lai (2006, 119) cite *Mencius* 4A17, as pointed out. The passage's focus on *quan* points out that, given the abnormality of the situation, the sister-in-law's being part of the family is a morally significant factor and thus the rule (*li*) of men and women not touching one another must give way to that factor or to rightness (*yi*). Concerning the seeming tension

between the three supporting papers' conception of Confucian ethics and Confucianism as skills-based, I point out that while it is true that Confucian ethics is multi-faceted and that the ethics has important internal debate, the idea that the ethics is skills-based does *not* exclude the socio-ethical or social roles in Confucianism. The proposal that it is skills-based is also primarily founded on the features of *junzi* in the *Analects* as discussed. *Junzi*'s features, as discussed by Cua (1971), point to the Confucian moral exemplar as a person of *ren* and *li*, which are highlighted in Wong. As presented, *junzi*'s features figure in the reasoning elements in Lai's discussion. Accordingly, it is not the case that the discussion of Confucian ethics as roles-based is unrelated to the idea that the ethics is based on skills of good ethical reasoning.

In this paper, I have argued that support for the claim that Confucianism is the best theory for AI and robotics is clear in three articles on comparative philosophy. The articles stress moral excellence (in Confucius's philosophy), which is one that emphasizes compassion (*ren*) and etiquette (*li*). Although Buddhist moral excellence is considered here, it is ultimately Confucianism's conception of moral excellence or a morally excellent person that is important. The conception leads back to the point of that conference on ethical and social concerns of AI in Asia (Checketts et al. 2025) that *xin* 信 (trustworthiness) ought to be used as the basis for AI design: "In designing the values of the AI, a thin Western concept of trustworthiness will be supplanted by the thicker, interrelational concept of *xin*." (Checketts et al. 2025, 5). According to the words, not only is the Confucian *xin* a specific, *traditionally*- or *culturally*-based concept, but it is characteristically also interrelational, meaning that unlike being trustworthy in common (Western) understanding or meaning (i.e., being *reliable*), reliability in *xin* is not egoistic or self-regarding. Confucian *xin* is an other-regarding virtue.[23] The morally excellent person in Confucianism is necessarily trustworthy because to be trustworthy is one important aspect and mark of being refined and being a person of *ren*. It is my suggestion here that refinement or moral excellence in the Confucian sense is, if not paramount, important to look into.[24]

NOTES

1. Although the term "Confucianism" is now used to denote the movement founded by Confucius, the movement or school is called *rujia* 儒家 (Bryan Van Norden, *Mengzi* (Hackett, 2008), 12). *Rujia* (school of the '*ru*'), however, is a term coined by the historian Sima Tan 司馬談 (d. 110 B.C.E.) (Kidder Smith, "Sima Tan and the Invention of Daoism, 'Legalism' et cetera," *Journal of Asian Studies* 62, no. 1 (2003): 129-56). It is also important to note that *ru* (儒) "predates Confucius, and connoted specialists in ritual and music, and later experts in Classical Studies" (Mark Csikszentmihalyi, "Confucius," *Stanford Encyclopedia of Philosophy* (2024), accessed June 18, 2025, at https://plato.stanford.edu/archives/sum2024/entries/confucius.

2. In arguing that even though quite a number of research studies have been done on possible contributions which Confucian ethics can give to virtue theory or virtue ethicists, the comparative work is still in its infancy, Justin Tiwald (2010) also points out detractors. Detractors compare Confucianism with the other two major

ethical theories of consequentialism and deontology. Tiwald also mentions the work of Roger Ames and Henry Rosemont Jr., who maintain that "the conceptual apparatus of Confucianism is fundamentally incompatible with an array of Western moral theories" (Justin Tiwald, "Confucianism and Virtue Ethics: Still a Fledgling in Chinese and Comparative Philosophy," *Comparative Philosophy* 1, no. 2 (2010): 58-9). Concerning care ethics and Confucianism, Li Chenyang (1994) claims that Confucian ethics is a care ethics. Daniel Star (2002) takes issue with Li's proposal, arguing that it is a virtue ethics.

3. That comparisons between these concerns and Confucian ethics have been made does appear clearly to indicate the multi-faceted nature of the ethics, or that the internal debate of Confucian ethics exists. However, given the features of *junzi* in the *Analects* that scholars have identified, I find Karyn Lai's discussion to be sound.

4. A note should also be made about the proposal of having that one theory as the solution to the problem. Given that the proposed pluralistic or multicultural approach to the problem seems the commonsensical solution, or given that any good ethics (theory) can be used to solve it, *why deal with the question?* Behind the pluralistic approach is the question, *Is there a way that the rather absolutist approach to the problem (due to imposition) be done away with?* Behind the argument that Confucian ethics or another should not be proposed, or it is futile to do so, because any good ethics or ethical theory can be used to solve it, is the question: *Does not the issue boil down to perspective?* One answer to the question of why consider the possibility is the soundness of the thinking that because all peoples now have to have guidelines on how to deal with the new and powerful technologies of AI and robotics, it is incumbent upon each to remodel their own cultural or religious traditions (for the simple reason that we might not be amenable to the values currently being seen as prominent for AI adoption or rejection) (cf. Soraj Hongladarom and Jerk Bandasak, "Non-Western AI Ethics Guidelines: Implications for Intercultural Ethics of Technology," *AI and Society* 39 (2024): 2028). The idea that we should *remodel* our tradition means that we must think that the values we hold are not irrelevant to facing the challenges of the present. Moreover, it appears permissible to hold the belief that our cultural, philosophical set of values might be the best for everyone. I believe that the possibility for the theory to be the norm exists because of everyone's *recognition* of its virtue and not somebody's or some culture's imposition.

5. Besides, in that the Buddhist theory's conception of a morally excellent person appears clearly to be comparable to Confucian *junzi*, the paradigmatic individual, this is a reason or starting point for saying that Confucianism can be the best theory.

6. Gabriel entertains the possibility of a machine that is able to discover the essence of morality. Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (2020): 425.

7. Hongladarom writes about the Trolley Problem for self-driving cars, probably because "the most famous ethical dilemmas for [these] is [sic] an adaptation of the trolley problem" (Soraj Hongladarom, *The Ethics of AI and Robotics: A Buddhist Viewpoint* (Rowman and Littlefield, 2020), 84). He ultimately points out that since, in one dilemma, a choice has to be made between putting the self-driving car's lone passenger safe by smashing into a bus load of forty young children (risking more

lives) or avoiding the bus and yet risking a life, it is incumbent upon designers to declare that choice *publicly* (Hongladarom, *Ethics of AI and Robotics*, 84-5. Lest no one buys self-driving cars programmed to choose the latter, Hongladarom states that what is important is that it is the public's task to debate over whether that programming ought to be legally banned or sanctioned, Hongladarom, *Ethics of AI and Robotics*, 85.

8. There are many interesting issues and points that Hongladarom discusses in the book. These include his discussions of (i) the question *Can robots be regarded as "persons"?*, (ii) the question of *machine enlightenment in autonomous technology,* (iii) machine enlightenment in *lethal autonomous weapons systems*, (iv) Shoshana Zuboff's stress on a "right to a future tense" (Hongladarom, *Ethics of AI and Robotics*, 146f) and (v) the possible provisions of AI for healthcare, education, and climate change.

9. This is Karyn Lai's translation. Karyn Lai, *Learning From Chinese Philosophies: Ethics of Interdependent and Contextualised Self* (Ashgate, 2006), 16, 61.

10. References to the *Mencius* (or *Mengzi*) text follow the concordance, ICS Mengzi Chapter/Page/Line number format. This can be directly input at https://ctext.org/mengzi or at https://ctext.org/tools/concordance#std1. *Chinese Text Project*, n.d. "Mengzi," accessed August 10, 2024, at https://ctext.org

11. References to the *Lunyu* (or the *Analects*) text follow the concordance, ICS Lunyu Chapter/Page/Line number format. This can be directly input at https://ctext.org/analects
or at https://ctext.org/tools/concordance#std1. *Chinese Text Project*, n.d. "Lunyu," accessed May 2, 2026, at https://ctext.org

12. "Etiquette," a translation stressed by Amy Olberding, is an important one. Amy Olberding, "Etiquette: A Confucian Contribution to Moral Philosophy," *Ethics* 126, no. 2 (2016): 422-446.

13. The similarity being drawn here between Buddhist *arahant* and Confucian *junzi* may be objected to as, at best, superficial, and that appealing to exemplification of compassion does not rule out many religious traditions, notably Christianity, in which compassion also plays a prominent role, and with a prominent role-model associated with compassion, viz. Jesus Christ. These are difficult to answer. The most that can be said here, I believe, is that the focus in this paper is on the possibility of Confucianism as the solution to the value alignment problem for AI and that this work, alas, is in the ultimate rather an exploration.

14. One very important objection in regard to the comparison, particularly between Confucianism and Buddhism, is that it does not touch on each of the traditions as a whole or in its entirety. According to this challenge, while there is a clear affinity between their conceptions of a moral exemplar, the comparison fails to consider important doctrinal differences between the two. Right at the "period of preparation," in Arthur F. Wright's discussion of Buddhism in Chinese history (see Arthur F. Wright, *Buddhism in Chinese History* (Stanford University Press, 1959), 38), apologetic writing was made. Wright writes: "Still another means of adapting and explaining Buddhism to the Chinese was apologetic writing. In such writing, generally there was a defense of the alien system, which not only extolled its merits but also pointed to ways in which it was either consonant with certain indigenous ideas and

values or complementary to them. An apologetic has a special value for the study of the interaction of two traditions because the points at which defense is felt to be necessary are invariably the points of greatest conflict between the two systems of ideas" (Wright, *Buddhism in Chinese History*, 38). Accordingly, that such was composed means that the traditions of Buddhism and Chinese (Confucian and Daoist) truly had ideas that are at odds. One example is Mouzi's 牟子 (3rd century C.E.) response, in the *Mouzi Lihuo lun* 牟子理惑論 ("Disposing of Error"), to the question of why Buddhist monks shave their heads, a practice incompatible with the Confucian teaching of taking care of, not injuring, one's body because they come from one's parents. In his response, Mouzi appeals to Confucius's praise of Tai Bo, who "cut his hair short and tattooed his body" and yet is one who exhibited "ultimate virtue" (Wm. Theodore de Bary and Irene Bloom, *Sources of Chinese Tradition, Vol. 1* (Columbia University Press, 1999), 423). I am grateful to Franklin Perkins for pointing this objection out to me.

15. It was held on October 16 and 17, 2024, at the Centre for Applied Ethics at Hong Kong Baptist University.

16. Pak-Hang Wong's works on Confucian philosophy of technology are cited as further resources in Version 2 of IEEE's *Ethically Aligned Design, which* includes the section "Classical Ethics". Section 2 of this part, "Classical Ethics from Globally Diverse Traditions" has sub-sections on Buddhism, Ubuntu philosophy, and Shintoism. The sub-section "The monopoly on ethics by Western ethical traditions" may perhaps be regarded as the Confucian section because it has the word "Confucianism," one of the only two pages in the document on which the word appears. But no Confucian ethical concepts or principles are mentioned or discussed. See Institute of Electrical and Electronics Engineers (IEEE), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version II* (2017), 203-11, accessed February 14, 2025, https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

17. Radical virtue ethics is the un-supplemented form, since this is the view that virtue ethical theory can stand on its own, without the aid or use of utilitarianism or deontology or other theories (James Rachels, *The Elements of Moral Philosophy*, 4th ed. (McGraw-Hill, 2003), 187-9). Wong's idea may be put as treating Confucian virtue ethics as radical virtue theory, because it is regarded as able to deal with *the just* and *the right*.

18. See footnote #30 on the page.

19. See also Antonio S. Cua, "The Ethical and Religious Dimensions of *Li* (Rites)," *Review of Metaphysics* 55, no. 3 (2002): 477-9, 481-2.

20. See *Mencius* 6A6.

21. Kremer, in his re-examination, mentions the United States' National Society of Professional Engineers (NSPE) Board of Ethical Review (BER) case no. 98-9. (Eugene Kremer, "(Re)examining the Citicorp Case: Ethical Paragon or Chimera?," *Arq: Architectural Research Quarterly* 6, no. 3 (2002): 274) The case is called "Duty To Report Unsafe Conditions/ Client Request For Secrecy" and features, as "facts," a scenario strikingly similar to the facts of Citicorp, where the questions under consideration are: (i) Is it ethical for Engineer A, the structural engineer, to comply

with the client's and the architect's desire for secrecy? (ii) Is it ethical for Engineer B, the city engineer, to maintain the secrecy? Case 98-9 is available at https://www.nspe.org/career-growth/ethics/board-ethical-review-cases/duty-report-unsafe-conditions-client-request

22. Avoidance of shame (*xiu* 羞) is also there but shame here relates more to (moral) *guilt*. It has been discussed that while *xiu* 羞 is frequently translated as "shame", it is not the same as *chi* 恥 (shame), because *xiu* 羞 in Mencius's philosophy "seems to refer to an internal moral-regulatory capacity" (Lai, *Learning From Chinese Philosophies*, 75). Accordingly, *xiu* is guilt. This is important for two reasons: (i) this points to the idea that Confucianism is not 'shame-morality', "whereby moral standards are mainly located in external sources" (Lai, *Learning From Chinese Philosophies*, 75 n. 30); (ii) this meaning appears to explain LeMessurier's action of presenting his experience or revealing the 'secret' two decades later.

23. The conference report points out that it may be difficult to imagine a trustworthy machine, and so design engineers surely need to be the ones who have *xin*. The report also says that these engineers ought to "balanc[e] the aims and needs of their firms with those of the users" (Checketts et al., "Ethical and Social Concerns of Artificial Intelligence in Asia," *Nanoethics* 19, no. 12 (2025): 5). This is *he* (harmony).

24. I thank the anonymous reviewers, special issue editors Professors Joseph Martin and Jose Ivan Efreaim Gozum, and the journal's editor-in-chief Professor Jove Jim Aguas, for the corrections and the invaluable comments and suggestions. I would also like to thank the University of San Carlos for the support through the RDEPO Research Grant for Semester 1, AY 2024-25.

REFERENCES

Chatila, Raja, Kay Firth-Butterflied, John C. Havens, and Konstantinos Karachalios. 2017. "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems." *IEEE Robotics and Automation Magazine* 24 (1): 110. https://doi.org/10.1109/MRA.2017.2670225

Checketts, Levi, Robert James Boyles, Benedict Shing Bun Chan et al. 2025. "Ethical and Social Concerns of Artificial Intelligence in Asia." *Nanoethics* 19 (12). https://doi.org/10.1007/s11569-025-00472-1

Chinese Text Project. n.d. "Mengzi." Accessed August 10, 2024, at https://ctext.org/mengzi/teng-wen-gong-ii

Csikszentmihalyi, Mark. 2024. "Confucius." *The Stanford Encyclopedia of Philosophy*. Accessed June 18, 2025, at https://plato.stanford.edu/archives/sum2024/entries/confucius/

Cua, Antonio S. 1971. "The Concept of Paradigmatic Individuals in the Ethics of Confucius." *Inquiry* 14: 41-55. https://doi.org/10.1080/00201747108601622

Cua, Antonio S. 2002. "The Ethical and Religious Dimensions of *Li* (Rites)." *Review of Metaphysics* 55 (3): 471-519. https://www.jstor.org/stable/20131749

de Bary, Wm. Theodore, and Irene Bloom, compilers. 1999. *Sources of Chinese Tradition: From Earliest Times to 1600, Volume 1*. 2nd ed. Columbia University Press.

Eno, Robert. 2016. *Mencius: Translation, Commentary and Notes*. Accessed August 10, 2024, at https://hdl.handle.net/2022/23423

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Mind and Machines* 30: 411-37. https://doi.org/10.1007/s11023-020-09539-2

Hauser, Larry. n.d. "Artificial Intelligence." *The Internet Encyclopedia of Philosophy*. Accessed August 18, 2024, at https://iep.utm.edu/artificial-intelligence/

Hongladarom, Soraj. 2020. *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Rowman and Littlefield.

Hongladarom, Soraj, and Jerd Bandasak. 2024. "Non-western AI Ethics Guidelines: Implications for Intercultural Ethics of Technology." *AI and Society* 39: 2019-32. https://doi.org/10.1007/s00146-023-01665-6

Institute of Electrical and Electronics Engineers (IEEE). 2017. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (Version II)*. Accessed February 14, 2025, at https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Jing, Shan, and Neelke Doorn. 2020. "Engineers' Moral Responsibility: A Confucian Perspective." *Science and Engineering Ethics* 26: 233-53. https://doi.org/10.1007/s11948-019-00093-4

Kremer, Eugene. 2002. "(Re)examining the Citicorp Case: Ethical Paragon or Chimera?" *Arq: Architectural Research Quarterly* 6 (3): 269-276. https://doi.org/10.1017/S1359135503001763

Kupperman, Joel. 1968. "Confucius and the Problem of Naturalness." *Philosophy East and West* 18 (3): 175-85. https://www.jstor.org/stable/1398259

Kupperman, Joel. 2002. "Naturalness Revisited: Why Western philosophers Should Study Confucius." In *Confucius and the Analects: New Essays*, edited by Bryan Van Norden. Oxford University Press.

Lai, Karyn. 2006. *Learning from Chinese Philosophies: Ethics of Interdependent and Contextualised Self*. Ashgate.

Lai, Karyn. 2012. "Knowing to Act in the Moment: Examples from Confucius' *Analects*." *Asian Philosophy* 22 (4): 347-64. https://doi.org/10.1080/09552367.2012.729324

Lau, D. C. 2003. *Mencius*. Penguin.

Legge, James. 1930. *The Four Books: Confucian Analects, the Great Learning, the Doctrine of the Mean, and the Works of Mencius*. The Commercial Press.

Leys, Simon. 1997. *The Analects of Confucius*. W.W. Norton and Company.

Li, Chenyang. 1994. "The Confucian Concept of *Jen* and the Feminist Ethics of Care: A Comparative Study." *Hypatia* 9 (1): 70-89. https://doi.org/10.1111/j.1527-2001.1994.tb00110.x

O'Leary, Daniel E. 2023. "An Analysis of Watson vs. BARD vs. ChatGPT: The Jeopardy! Challenge." *AI Magazine* 44: 282-95. https://doi.org/10.1002/aaai.12118

Olberding, Amy. 2016. "Etiquette: A Confucian Contribution to Moral Philosophy." *Ethics* 126 (2): 422-446. https://doi.org/10.1086/683538

Rachels, James. 2003. *The Elements of Moral Philosophy*. 4th ed. McGraw-Hill.

Smith, Kidder. 2003. "Sima Tan and the Invention of Daoism, 'Legalism,' et cetera." *Journal of Asian Studies* 62 (1): 129-56. https://doi.org/10.2307/3096138

Star, Daniel. 2002. "Do Confucians Really Care? A Defense of the Distinctiveness of Care Ethics: A Reply to Chenyang Li." *Hypatia* 17 (1): 77-106. https://doi.org/10.1111/j.1527-2001.2002.tb00681.x

Tiwald, Justin. 2010. "Confucianism and Virtue Ethics: Still a Fledgling in Chinese and Comparative Philosophy." *Comparative Philosophy* 1 (2): 53-63. https://doi.org/10.31979/2151-6014(2010).010207

Van Norden, Bryan. 2002. "Introduction." In *Confucius and the Analects: New Essays*, edited by Bryan Van Norden. Oxford University Press.

Van Norden, Bryan. 2008. *Mengzi*. Hackett.

Wong, Pak-Hang. 2012. "*Dao*, Harmony and Personhood: Towards a Confucian Ethics of Technology." *Philosophy and Technology* 25: 67-86. https://doi.org/10.1007/s13347-011-0021-z

Wong, Pak-Hang. 2020. "Why Confucianism Matters for the Ethics of Technology." In *The Oxford Handbook of Philosophy of Technology*, edited by S. Vallor. Oxford University Press.

Wright, Arthur F. 1959. *Buddhism in Chinese History*. Stanford University Press.