**How AI Trained on the Confucian *Analects* Can Solve Ethical Dilemmas**

The influence of AI has spread globally, intriguing both the East and the West. As a result, some Chinese scholars have explored how AI and Chinese philosophy can be examined together, and have offered some unique insights into AI from a Chinese philosophical perspective. Similarly, we investigate how the two fields can be developed in conjunction, focusing on the popular Confucian philosophy. In this work, we use Confucianism as a philosophical foundation to investigate human-technology relations closely, proposing that a Confucian-imbued AI would reduce prevalent AI ethical issues. Through this discussion, the strong substantivist stance that we will lose control over AI is refuted. We then discuss the practical elements of this project, explaining how passages from the Confucian text of *the Analects* can be incorporated into AI systems. As an example of a real-world context, psychology is used to demonstrate how the project could be practically applied. Potential criticisms surrounding the proposed project are also identified and challenged. By reducing the negative effects of common AI ethical issues and utilising Confucian principles from *The Analects*, AI systems can be more ethical, promoting peaceful relations between humans and technology.

**1 Introduction**

Artificial intelligence (AI) is without a doubt showing exponential growth every day. It is a field that explores the way in which machines can replicate humans’ ability to perform different tasks, such as decision-making, translation, driving, and understanding people. Unlike traditional computational systems, AI systems can learn and improve based on historical data. They mainly involve algorithms, which are developed via the artificial neural network (ANN).

The ANN is specifically modelled to replicate the electrical activity within the human brain and nervous system. Neurodes, which are typically arranged in layers or vectors, are connected to other neurodes. Weighted data signals then enter a neurode to transfer the information within the network. Since AI systems contain the ANN, they are able to process information similar to how human brains can, thereby enabling them to self-learn (Walczak, 2003).

Seeing that AI is a complex recent technological advancement that functions in a way that is similar to humans, investigating its potential societal impacts, especially relating to its ethical nature, is of utmost importance. In light of this, the proposed project puts forward a novel idea of technological development within the field of AI and computer science: inserting passages from *the Analects*, the most well-known piece of literature by ancient Chinese philosopher Confucius into the algorithms of AI. *The Analects*, based on Confucius’ teachings to his students, is a seminal work on virtue ethics that has been used for centuries to cultivate a holistic sense of virtue in individuals, and its moral principles are often considered the cornerstone of modern Chinese values.

There are a few main reasons as to why *the Analects* is chosen to be used in this proposed technological advancement. One of the main reasons is the fact that virtue ethics, a moral philosophy that stresses the importance of building a moral character through the cultivation of virtues, although researched by some, is not an area of philosophy that is as thoroughly developed in Western realms, such as deontology. Another reason is that there is a mounting number of concerns over the high level of eurocentric biases in the data of AI systems, so using an eastern philosophy would help reduce them. Given that Confucianism is the most well-known Chinese philosophy in modern-day society and had widespread influence in feudal China, there is a natural desire to investigate whether it could still be applicable to contemporary society. A third reason as to why *the Analects* is chosen is that the entirety of it is written in aphorisms, providing its readers with valuable insights into what it means to be virtuous, according to Confucianism. Aphorisms, unlike traditional full-picture stories, are not often examined closely, which is why this concept for a technological innovation seeks to investigate this area, as well as to defend their significance.

It is therefore with confidence that this proposal for a technological innovation is put forward in the field of AI along with high hopes that this postphenomenological investigation will contribute to the advancement of human society.

**2. Confucianism and AI Ethical Concerns**

Currently, there are many ethical concerns within the field of AI, with the most common ones being: (1) that AI will replace human jobs; (2) that AI will develop in ways that cannot be controlled by humans; (3) that there are biases in the data used to train the AI models; and (4) that AI systems are not private or secure enough.

 Although the concept of combining Confucianism and AI systems is relatively new, there is still some existing research that supports the idea. This means that using a Confucian approach to alleviate AI ethical concerns can be seen as suitable. For example, philosopher and ethicist of technology Dr. Pak-Hang Wong demonstrates the importance of the Confucian virtue of *li* (禮, propriety) through his paper *Why Confucianism Matters for the Ethics of Technology* (2020), emphasising that *li* is important to the ethics of technology for primarily two reasons: (1) that it makes up the rules for social interaction, similar to how grammar makes up the rules for languages; and (2) that it fully embodies the virtues of the Confucian tradition and that it provides a set of moral or ethical principles that are considered binding within that tradition, which Wong refers to as the “normative standard”. Ultimately, he argues that *li* can serve as the “normative standard” for ethical considerations, design, and use, or the ritualisation of technology to guide users and societal responses with reference to the virtue.

Additionally, in *Intelligence and Wisdom* (Song, 2020), contributor Chengyang Li also recommends AI to be included in the Confucian moral domain, which includes people and animals, which exhibit intelligent signs of morality. Some scholars believe that the intellect of AI systems means it should be included. Others cite the Confucian order, arguing that AI’s status as a *wu* (inanimate object, 物) disqualifies it, as *wu* is the lowest category in the moral hierarchy. However, Li reasons that despite this it should be “upgraded” within the Confucian order if it is able to make ethically relevant decisions, making it functionally similar to people and animals.

With the support of existing research within the Confucianism and AI discipline, there is promise that integrating passages from *the Analects* into AI systems would reduce AI ethical issues as Confucian ethical principles and values promote fairness, equity, and inclusivity.

**2.1 The concern that AI systems will eliminate human jobs**

The concern that AI systems will replace human jobs arises from a lack of understanding or misconceptions regarding the advancements, capabilities, and limitations of AI technology. The fear that AI systems will replace human jobs can be broken down into different “subtypes”, each of which adds to the overall concern about the potential elimination of human jobs by AI systems. The first “subtype” is the fear that smart machines can outthink us, therefore having the ability to replace human judgement.

The question of whether or not AI could replace human jobs from a decision-making aspect was explored by a group of Harvard professionals. One of the professionals, Fuller, responds that it is highly unlikely for AI to eliminate entire job categories. He states that “what we’re going to see is jobs that require human interaction, empathy, that require applying judgement to what the machine is creating [will] have robustness” (Pazzanese, 2020). Another professional, Mills, believes that AI could bring more benefits than harm to businesses because their owners could be provided with detailed insights into important financial information pertaining to them (Pazzanese, 2020). Therefore, as long as AI is used as a supporting device instead of a primary method for humans to make decisions, it will only enhance their ability to make informed ethical judgments and will *not* eliminate human jobs as a result.

This aligns with the Confucian belief that education and self-cultivation are necessary for moral development, which is elucidated in *the Analects* 2.15: “The Master said, “to study without thinking will cause bewilderment. To think without studying will cause danger” (「學而不思則罔，思而不學則殆。」) (Lui, 2016). Contextualising Confucius’ philosophy, it means that humans would be foolish not to *think* or *study* the implications involved in their ethical decisions before using AI to aid them in their decision making. This is also in agreement with Naik et al.’s paper, which argues that users or operators of AI systems should be aware of their limitations and how they are designed, even when bias within the systems are effectively managed (2022). Whilst AI can enhance human decision-making, humans should still remain educated about the capabilities and limitations of AI and use it responsibly.

**2.2 The concern that humans will lose control over AI.**

The second “subtype” of fear is the concern that the power of AI will progressively surpass human control. This can be addressed via the study of postphenomenology, which aims to explore the way in which technologies help to shape the relations between humans and the world. The proposed project, that passages from *the Analects* should be incorporated into AI systems, can contribute to postphenomenolological studies by providing a philosophical foundation that explores the relationship between humans and technology. Through this, humans are invited to reflect on the role of technology and to consider whether technology is fundamentally an instrument that derives all its impact from human actions and intentions, or if it possesses the capacity to reshape society beyond, and potentially independent of, human intentions (Rosenberger and Verbeek, 2015, p.2). The former approach is known as instrumental- ism, whereas the latter approach describes a strong substantivist stance. Not *all* substantivists view technology as pessimistically as described. In fact, only very few support such a strong stance. Milder substantivists simply acknowledge the non-neutrality of technology as a tool (Rosenberger and Verbeek, 2015, pp.2-3).

The belief that AI will surpass human control and replace human jobs is a thought that is driven by certain sectors of substantivists, which the proposed project seeks to address. The most prominent example of a strong substantivist is German philosopher Heidegger, who views technology as “the greatest danger”. He believes that danger is in itself dangerous, and that humans have not unravelled its true form or essence. Hence, he reasons that humans should develop a mindset of “releasement”, which involves recognising that technology is omnipotent and that it is impossible to escape from understanding the world under its influence. He explains that instead of succumbing to it, humans should instead stay open to new ways of understanding and existing in the world (Heidegger, 1977).

There are some problems with regards to his approach. The first problem, which is identified by political philosopher Feenberg (2000), is that it fails to provide scholars who want to engage with differing technologies and their societal implications with anything of merit, since it assumes that all technology is harmful to the same extent. A second limitation outlined by Verbeek (2005) states that Heidegger’s approach is limited, as it does not consider the practical and real-world implications of technology. A third limitation is the overemphasis on the historical development of humans and the role that technology plays within it. Even though it is important to consider the historical nature of humans, the evolving nature of technology should not be undermined.

It can therefore be argued that the strong substantivist stance is highly limited, further reflecting the need for humans to view technology more open-mindedly. Humans should, for instance, be encouraged to reevaluate their fears pertaining to AI eliminating human jobs. The proposed project supports this, resonating with the broader idea of postphenomenology, promoting that humans can work alongside technology. It seeks to maximise the benefits of AI whilst ensuring that human control remains the dominant force behind the development of the systems. Additionally, it adopts an instrumentalist stance, acknowledging that AI is neutral in itself- the implications of technology depends on how it is designed by, and how it interacts with, humans. Ergo, the project seeks to address the concern that AI will surpass human control and replace jobs by directly contributing to the growing field of postphenomenology because it will examine the relationship between AI systems and their users closely.

**2.3 The concern that there are biases in training data.**

The concern that there are biases in the data used to train AI models can be alleviated with the proposed project. Training data is an essential and arguably the most important part of machine learning. It is a set of data that the machines use to correct their behaviour and is the foundation of the machine learning model, which corrects the behaviour of the machine through analysing the training data and understanding its features (Joby, 2021).

Different factors that contribute towards the existence of biases in training data include the underrepresentation or overrepresentation of certain demographic groups, sampling biases, data collection methods, and inherent biases in the data sources themselves (Chapman University, 2022).

The proposed project will target the factor of the underrepresentation of data due to the diverse perspectives offered by Confucius’ teachings from the Analects regarding the various aspects of human life, focusing on the importance of virtues. By integrating the Analects’ passages into AI systems, the models are exposed to a wider range of ethical viewpoints, challenging the limitations of Western-centric biases that may be present in the data. This infusion of diverse perspectives allows AI models to develop a more inclusive understanding of human values and behaviours. A way of approaching this is through the study of virtue ethics. This is better understood as behaviours of high moral standards, so that the right decision can be made in any given circumstance (Angle and Slote, 2013).

Namely, some important virtues of Confucian philosophy are ren (仁), the concept of humaneness; xiao (孝), which is filial piety; li (禮 ), meaning propriety; and yi (義), which refers to righteousness.

The core of Confucian philosophy, however, is the virtue of ren (仁). It describes that people should behave in a loving, selfless, and moral way. Confucius believed that through building a moral character, it enables one to affect the world around them positively through “cosmic harmony”, which refers to the maintenance and regulation of various properties in the world. It makes sense, then, that the golden rule of Confucianism is “do not do unto others what you would not want others to do unto you” (Angle and Slote, 2013).

It is through *ren* (仁) that the other virtues are introduced: *xiao* (孝) describes the way in which people should interact with each other based on their relationship status, *li* (禮) concentrates on the fixed set of rites and rituals that people should adhere to according to the role system, and *yi* (義) is the willingness to behave in a positive way that is beneficial to oneself and the community around them. *Xiao* (孝), for instance, would be for one to treat their parents with respect; *li* (禮)would be for a funeral guest to wear black to the funeral; and *yi* (義) would be for one to volunteer for military service.

By taking advantage of the emphasis on virtue ethics that is evident through the Analects, the biases in training data of AI systems will be challenged through a more holistic and morally grounded approach to artificial intelligence development and implementation.

**2.4 The concern that there are privacy and security issues**

The virtue of *xiao* (孝) can serve as a guiding principle for the development of AI systems, addressing concerns about privacy and security. By inserting the relevant passages related to *xiao* (孝) from the Analects into the AI systems, the privacy and security of users can be prioritised.

This is because if AI systems are imbued with *xiao* (孝), they will be programmed to act with respect towards humans, meaning that they will ensure that user data is handled with care and not misused or accessed without consent.

The belief that all relationships must fall under a strict social hierarchy is closely related to *xiao* (孝), as it emphasises the importance of respecting and fulfilling one's obligations to family members and other individuals based on their social roles and positions. This can be seen through the *five cardinal relations* (五倫), which refer to: (1) the relationship between sovereign and subject; (2) the relationship between father and son; (3) the relationship between elder and younger brother; (4) the relationship between husband and wife; and (5) the relationship between a friend and friend (He and Li, 2015).

According to Confucian philosophy, family lies at the centre of society and should be prioritised before the individual (Stanford Medicine, 2019). A passage from *the Analects* that illustrates this is 4.18, in which “The Master says, “in serving our parents, if we discover some fault in them, we should advise them tactfully by expressing our own view. If they do not take any heed, we should still respect them rather than rebel against them. While we worry about them, we should not blame them” (子曰：「事父母幾諫。見志不從，又敬不違，勞而不怨。」) (Lui, 2016).

This can be used to address privacy issues surrounding AI because it mentions the necessity of respect and empathy within a parent-child relationship. This means that if it were to be incorporated within AI systems, a sense of ethical behaviour and responsibility could be instilled by the developers, ensuring that user data is handled with care and respect. As a result, the users will feel more safe and secure in handling AI systems.

**3 The Method of Integration of *the Analects***

Machine consciousness, as a hot topic in contemporary society, will inevitably arise as a discussion point when considering this project. Would the project require machines to be conscious?

The idea of machine consciousness has also been discussed amongst Confucian scholars. Contributor Chunsong Gan of *Intelligence and Wisdom* (Song, 2020) is pessimistic about this, fearing the impacts that it might have on traditional reproductive practices and the Confucian ideal of “kinship society”.

Despite the instrumentalist stance of the proposed project, it will still not support the idea of machine consciousness, as the repercussions and consequences of such an approach could be tremendous. In order for machine consciousness to be considered, more research with regards to it is required.

Also, machine consciousness is not needed for this proposed project in particular, as it could integrate Confucian ethics into AI by incorporating explicit ethical maxims into the machine's code, enabling precise ethical judgments based on Confucian principles in different ethical situations. This is explored through James H. Moor’s paper on *The Nature, Importance, and Difficulty of Machine Ethics* (2006), in which he describes two ways of incorporating ethics into machines: the first way is to create software that implicitly supports ethical behaviour (which is what implicit ethical agents entail), and the second way is the aforementioned one, which is to write code containing explicit ethical maxims.

Through the first way, Moor explains that the machine’s actions will be constrained and so unethical outcomes will be avoided, and that it will also act ethically because the machine’s internal functions will promote ethical behaviour. Moor also argues that in order for machines to achieve ethical outcomes, it is more feasible for them to act as implicit ethical agents. He uses the implicit ethical agents of automated teller machines and Web banking software to illustrate his point- they are meticulously constructed, and argues that using a line of code to tell the machines to be honest will not produce the desired outcome.

AI systems are implicit ethical agents and therefore the proposed project would adopt this approach. In the context of this project, the AI would be trained to act according to Confucian principles in numerous situations. Over time, it would learn to internalise the behaviour and act accordingly in unseen situations. For example, if someone were to ask the AI if they should hire someone who has flattered them during an interview by referring to them as “top of the field” despite the candidate lacking experience and showing no signs of a strong work ethic, the AI could be taught to respond according to the passages from the first chapter of *the Analects*, which is about learning. In this situation, for instance, it could respond: “The Master says, “Those who are glib in their speech and wear an ingratiating expression have little benevolence about them” (「巧言令色，鮮矣仁」) (Lui, 2016). The response could also include a contemporary explanation of the saying in the context of the situation. Below the saying, it could state: “be careful with flattery. Flattery is not always the truth.”

If this project were not AI-based and instead adopted an explicit agents approach, the number of rules required to be inserted into the system would be inexhaustible. Due to the contingent nature of virtues found in *the Analects*, there are many possibilities for each virtue, meaning that it would take immense periods of time and effort to ensure that each possibility is inserted into the system. In addition, the system would be limited to only a few situations, as it would only know how to respond in situations that it has been taught. In comparison, an implicit ethical agent approach would provide us with more benefits.

Machine consciousness exists in theory, whereas an explicit ethical agent approach is not feasible for the purpose of this project. An implicit ethical agent approach, on the other hand, would be an efficient and energy-saving way of achieving the goal of having a Confucian-imbued AI that is trained on *the Analects.*

**4. The Practical Application of The Project Explored Through Psychology**

Inserting passages from *the Analects* into AI systems could benefit various industries in a number of ways, as it promotes ethical behaviour that should be taken into consideration no matter which industry one may work in.

*The Analects* contains values that align with the moral beliefs of a wide variety of people from differing backgrounds, emphasising the concepts of trust and kindness, meaning that an AI that is embedded with Confucian principles could benefit an unprecedented number of people.

The field of psychology witnesses expansion on a daily basis, especially within the area of psychotherapy, following a societal paradigm shift that places a strong emphasis on mental health in recent years. Since it is a discipline that explores interpersonal relations closely, it is reasonable to evaluate how the proposed technological advancement could be implemented within this industry.

The concept of AI chatbots functioning as therapists, although new, is rapidly advancing, where an increasing number of experts in both the fields of psychology and computer science are exploring ways to use AI as a tool for counselling or therapy. Examples of this include Woebot (Evers, *et al*., 2024), talk2us.ai (talk2us.ai, 2024), and Ellipsis Health (Verma, *et al*., 2016), which are all AI chatbots that use cognitive behavioural therapy (CBT) principles to provide users with guidance and emotional support pertaining to each of their respective situations. Out of the three, Ellipsis Health can even detect in advance warning signs of mental disorders, such as anxiety and depression, before they fully manifest (Verma, *et al*., 2016). As a result, users are able to more proactively change their daily lives before their issues develop chronically.

Even though these chatbots are relatively safe to use, the topic of mental health issues is a sensitive topic and can even lead to health and safety concerns. Before the user interacts with the chatbot, therefore, these platforms usually offer a disclaimer stating that the company does not guarantee the full accuracy of information provided by the chatbot, and recommends one in crisis to seek help from a nearby emergency room or online hotlines. With this emerges the issue of ethics: to what extent are AI chatbots ethical?

The American Psychological Association (APA) keenly distinguishes between the Ethical Standards of psychology and its General Principles (American Psychological Association, 2017). The Ethical Standards of psychology include guidelines on resolving ethical issues, competence, human relations, privacy and confidentiality, and more.

General Principles, on the other hand, suggest ethical ideals that are only “aspirational in nature” (American Psychological Association, 2017) to encourage the highest order of ethical standards. They include ethical recommendations such as beneficence and nonmaleficence, fidelity and responsibility, integrity, justice, and respect for human rights and dignity.

These principles align with the Confucian principles that are illustrated in *the Analects*. For example, regarding beneficence, a passage from *the Analects* that depicts this is Chapter 4.2, where The Master says, “Only those who are wise and benevolent can seek humanity” (「仁者安仁，知者利仁」。) (Lui, 2016). Another example is Chapter 4.19, where The Master says, “When your parents are around, do not take a long journey. And if you must do so, you should let them know where you are heading.” (「父母在，不遠遊。游必有方」。) (Lui, 2016). This aligns with the ethical recommendation of responsibility in the General Principles of Psychologists and Code of Conduct, as it promotes responsible behaviour by advising us to inform our parents of our location to prevent them from getting worried. This means that the proposed project can not only enhance the Ethical Standards set by the APA within psychology AI chatbots, but also substantialise the General Principles, promoting them from aspirational principles to compulsory ones.

Between the APA’s Ethical Standards and *the Analects*, an overlapping feature that they share is the concept of benevolence, which is referred to as *ren* (仁) in *the Analects*, and Beneficence and Nonmalevolence in the APA Ethical Principles of Psychologists and Code of Conduct (2017). According to APA’s definition of Beneficence and Nonmalevolence, it is when psychologists do no harm to their clients and instead strive to benefit them.

This aligns with what is elucidated in *the Analects*, as Confucius states that it is important to “be loyal, honest and reliable when dealing with people '' when Fan Chi asks about benevolence in 13.19 of *the Analects*. This means that treating others well is a recurring theme in both Confucian philosophy and ethical guidelines for psychologists. Hence, if the passages from *the Analects* related to benevolence were to be incorporated into the algorithm of psychology AI chatbots, they would be able to respond in a manner that is in line with the Ethical Standards whilst emphasising Confucian teachings at the same time. For example, in a hypothetical situation, if the user of the Confucian-imbued chatbot were to ask the AI if they should continue living with their toxic partner or to move out and live with their kind sister instead, it could then respond: “according to Confucius, “living in an environment of benevolence is beneficial” (“里仁为美，则不处仁”).

Another overlapping principle between the APA’s Ethical Standards and *the Analects* is the concept of trust. This is of paramount importance regarding AI chatbots and user interactions, as users wish to disclose sensitive information to the chatbots, which can only be done under the condition that the users trust the chatbots. According to Confucius, “if a person cannot keep his word, he will not be able to achieve anything” (“人而無信，不知其可也”). In a hypothetical situation, if a user of a Confucian-imbued AI chatbot were to inquire about their partners’ infidelity, the chatbots would be able to give advice to them according to Confucius’ sayings. It could then also offer a contemporary analysis of the situation and contextualise it to fit the user’s needs. In this circumstance, the AI could, for instance, respond: “According to Confucius, “人而无信，不知其可也。” This translates to “if a person cannot keep his word, he will not be able to achieve anything.” Therefore, you should believe that he is keeping his word in saying that he will be bound to you forever- for if it is trust that he lacks, he may not be able to achieve anything of significance in your relationship.”

**5. Counterarguments to the Proposed Project**

However, despite the many benefits of the proposed project, some factors may bring into question its validity. These counter arguments will be addressed and discussed at length, including the aphoristic nature of *the Analects* and critiques from Daoists. It will be explained that using the aphorisms from *the Analects* do not limit the data that the AI receives and instead opens the system to new possibilities. It will also be explained that this approach would not lead us astray from nature and will instead bring us closer to The Way, otherwise known as *dao* (道).

**5.1 The aphoristic nature of *The Analects***

The proposed technological advancement, that passages from the Analects should be inserted

into AI algorithms, is centred around the usage of *the Analects*. Although some may criticise the use of an aphoristic piece instead of a traditional extended argument or essay, aphorisms can be perceived as more preferable because they are subject to contingency in a way that is within reason. This is supported by philosopher Wong (2020), as he explains in his paper that the Confucian ideal of “naturalness” (自然) enables humans to reflect on not only *what* we do in a situation, but also *how* we do it. It is a *pattern*, as it is consistent throughout *the Analects*, no matter which virtue is emphasised. For example, 13.23 of *the Analects* states that “The noble man is in harmony but does not follow the crowd. The inferior man follows the crowd, but is not in harmony” (“君子和而不同。小人同而不和”) (Muller, 2021). This accentuates the underlying importance of naturalness because it is implied that if one were to actively change their behaviour to fit in with others, their behaviour would not be authentic as it is forced. Everyone has a different “naturalness” which should be accepted. Philosopher Kupperman (2002) further elaborates on this idea, as he explains in his paper that the Confucian ideal of “naturalness” is when “the agent is reasonably comfortable with his or her behaviour, and there is no conflict between the behaviour and what the agent normally is like”, which is in alignment with what Wong illustrates.

The contingent nature of the virtues can also be found throughout other parts of the Analects. For example, both 12.22 and 12.1 of The Ana*lects* state the word “*ren*” (仁), but their definitions in each situation differ due to them being used in different contexts. For example, in 12.22, when Fan Chi, one of Confucius’ disciples, asks Confucius for the meaning of *ren* (仁), he responds with “love your fellow men” (“爱人”). However, when Yan Yuan poses the same question in 12.1, Confucius responds “benevolence is to discipline yourself to complete tasks according to the protocol… You should see no evil, hear no evil, say no evil, and do no evil.” It is therefore evident that the virtues are purposely subject to contingency, and that this is an essential element that should not be carelessly misplaced or overlooked when extracting Confucius’ ideas from *the Analects* to use for another project.

**5.2 A Daoist critique**

Another possible piece of criticism regarding the project might come from the Daoists, who are philosophers that believe in embracing *dao*, otherwise known as “The Way” (道), and maintaining “Oneness” (守一). They might criticise the proposed project for disrupting nature through promoting the use of technology. They believe that it would interrupt “The Way”, which is the universal operating mechanism of the world. As Song (2020) comments in *Intelligence and Wisdom*, “they are unlikely to embrace disruptive technologies that result in any displacement of the natural flow of things.”

However, it can be argued that technologies can bring humans closer to *dao* (道), especially since postphenomenological investigations place a particular emphasis on how humans and technology can interact with each other in a mutually beneficial way. As Verbeek (2021) illustrates in his paper, these studies shed light on the relations between humans as they are interacting with different types of technology, so that these technologies serve as a mediation tool instead of a disruptive one. Examples of this include using a MRI scanner, or an EEG scanner to plan a patient’s treatment plan, where humans are directly interacting with the technology and develop new types of behaviour as a result of their interactions. Furthermore, instead of viewing technology as a foreign entity that disrupts the flow of nature, it is also worth considering that it is now seamlessly embedded into our daily lives, with its presence now feeling as natural as any other element of our environment. Hence, it might be worth considering the possibility that technology might be now essential to maintaining the natural flow of life in contemporary society. By viewing technology through a postphenomenological lens and considering the positive impacts of technologies, the Daoist critique can be refuted.

**6. Conclusion**

The development of AI has a variety of ethical implications, pervading both interpersonal interactions and the mechanisms of our societies. An over-reliance on AI raises many concerns regarding the possibility of AI replacing human jobs, biases in the data that is used to train AI models, and that the AI systems that we use in our daily lives are not secure enough. This project therefore proposes a technological advancement in the field of AI- or, more specifically, AI ethics. Namely, the proposal is to insert passages from the Analects, an important piece of literature in Ancient Chinese history created by philosopher Confucius’ disciples, into AI algorithms to promote more ethical AI-powered systems.

Not only does the proposed technological advancement acknowledge ethical issues within the field of AI and offer ways to reduce them, but it also introduces postphenomenology as a recent field that is being investigated by computer scientists and philosophers alike. It defends the advancement of technology in the sense that it can be used as a source of comfort when considering the instrumentalist stance within the field of postphenomenology. Moreover, since it is proven that AI developers and philosophers are investigating relations between humans and technology, meaning that there is a growing awareness of AI ethical issues, and increasing attempts to alleviate them. The idea of a Confucian-infused AI, then, seeks to contribute to postphenomenological studies.

Additionally, to illustrate that technology can ethically play a role in humans’ lives through the proposed technological advancement, this dissertation uses the field of psychology as an example of an industry that could benefit from it. Since it is a discipline that deals with personal and sensitive issues, it is then believed that the Confucian-imbued AI could be applied to a wide range of industries.

Lastly, this project addresses the potential criticisms that it may receive due to its bold, original, and modern approach. It understands the complexities involved pertaining to the interpretational aspect of the proposed development, and also argues for the nature of aphorisms as a strength instead of a weakness.

As AI inevitably continues to evolve, humanity must continue to develop it with rigour and careful consideration of ethics, involving ethical philosophers and professionals in its field of application at every step to navigate its benefits and limitations.

**Author Contributions** Author E.S fully contributed to this paper.

**Funding** No funding was received to assist with the preparation of this manuscript.

**Data availability** Not applicable.

**Code availability** Not applicable.

**Declarations**

**Conflict of interest**  None to declare.

**Consent to Participate** Not applicable.

**Consent for Publication**  Not applicable.

**Ethical Approval** Not applicable.

**References**

Acmuller.net. (2021). Analects of Confucius 論語. [online] Available at: https://www.acmuller.

net/con-dao/analects.html

American Psychological Association (2017). ETHICAL PRINCIPLES OF PSYCHOLOGISTS

AND CODE OF CONDUCT. [online] American Psychological Association. Available at:https://www.apa.org/ethics/code/ethics-code-2017.pdf

Evers, M., Darcy, A., and Levy, M. (2024). Woebot Health. [online] Available at: https://woebot

health.com/ our-leadership/

Feenberg, A. (2000). From essentialism to constructivism: Philosophy of technology at the

crossroads. [online] ResearchGate. Available at: https://www.researchgate.net/ publication/260983580\_From\_essentialism\_to\_constructivism\_Philosophy\_of\_technology\_at\_the\_crossroads

Heidegger, M. (1977). The Question Concerning Difficulty of Machine Ethics. Technology and

Other Essays. New York, USA: Garland Publishing, Inc.

He, H., and Li, C. (2015). *Social Ethics in a Changing China: Moral Decay or Ethical*

*Awakening? JSTOR*. [online] Available at: https://www.jstor.org/stable/10.7864/ j.ctt7zsw42

Joby, A. (2021). What Is Training Data? How It’s Used in Machine Learning. [online] G2.com.

Available at: https://learn.g2.com/training-data

Kupperman, J. (2002). Naturalness Revisited: Why Western Philosophers Should Study

Confucius. [online] pp.39–52. doi:https://doi.org/10.1093/oso/9780195133950.003.0002

Lui, T. (2016). The Analects: A Modern Translation and Contemporary Interpretation, 論語今譯

時析. 2nd edn. Edited by Wong,W.K and Cheng, T: Yew Chung Publishing House.

Moor, J.H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent

Systems, [online] 21(4), pp.18–21. doi:https://10.1109/ mis.2006.80

Naik, N., Hameed, Z., Shetty, D.K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S.,

Patil, V., Komal S., Shetty, S., Bhavan P.R, Piotr C. and Somani, B.K. (2022). Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?, [online]. doi:https://doi.org/10.3389/fsurg. 2022.862322

Pazzanese, C. (2020). Ethical concerns mount as AI takes bigger decision- making role. [online]

Harvard Gazette. Available at: https://news.harvard.edu/gazette/story/2020/10/ethical- concerns-mount-as-ai-takes-bigger-decision-making-role/

Rosenberger, R & Verbeek, PPCC (ed.) 2015, Postphenomenological Investigations: Essays on

Human-Technology Relations. Postphenom- enology and the Philosophy of Technology, Lexington Books, Lanham, MD

Song, Bing (ed.), Intelligence and Wisdom, CITIC Press Group, 2020.

Stanford Medicine (2019). Cultural Beliefs and Practices: Role of Confucianism. [online]

Available at: https:// geriatrics.stanford.edu/ethnomed/chinese/fund/beliefs.html

Stephen C. Angle and Michael Slote (eds.), Virtue Ethics and Confucianism, Routledge, 2013.

Talk2us.ai. (2023). [online] talk2us.ai. Available at: https://www.talk2us.ai/

Verbeek, PP. 2005. What Things Do: Philosophical Reflections on Technology, Agency, and

Design. University Park, PA: Penn State University Press.

Verma, A., Lansinger, B., Collins, E., Mondal, M., McCool, M., Aratow, M., and Solinsky, S.

(2016). Ellipsis Health. [online] Available at: https://www.ellipsishealth.com/ about-us

Walczak, S. (2003). Artificial Neural Networks. Encyclopedia of Physical Science and

Technology, [online] pp.631–645. Available at: https://www.academia.edu/157 26358/Artificial\_Neural\_Networks

Wong, P.H. (2020). Why Confucianism Matters for the Ethics of Technology, [online]

pp.608-628. doi: https://doi.org/10.1093/oxfordhb/978019085 1187.013.36